

**PCT**WORLD INTELLECTUAL PROPERTY ORGANIZATION  
International Bureau

## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

<b>(51) International Patent Classification <sup>6</sup> :</b> <b>C12N 15/12, C12Q 1/68, C07K 14/47,</b> <b>16/18, A61K 38/17</b>	<b>A2</b>	<b>(11) International Publication Number:</b> <b>WO 99/67384</b> <b>(43) International Publication Date:</b> 29 December 1999 (29.12.99)
<b>(21) International Application Number:</b> PCT/US99/13524 <b>(22) International Filing Date:</b> 15 June 1999 (15.06.99) <b>(30) Priority Data:</b> 09/102,615 22 June 1998 (22.06.98) US <b>(63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Application</b> US 09/102,615 (CIP) Filed on 22 June 1998 (22.06.98) <b>(71) Applicant (for all designated States except US):</b> INCYTE PHARMACEUTICALS, INC. [US/US]; 3174 Porter Drive, Palo Alto, CA 94304 (US). <b>(72) Inventors; and</b> <b>(75) Inventors/Applicants (for US only):</b> WALKER, Michael, G. [CA/US]; Unit 80, 1050 Borregas Avenue, Sunnyvale, CA 94089 (US). VOLLMUTH, Wayne [US/US]; 783 Robie Avenue #1, Menlo Park, CA 94025 (US). KLINGLER, Tod, M. [US/US]; 28 Dover Court, San Carlos, CA 94070 (US). SPRINZAK, Einat, A. [IL/IL]; 20 Taraad Street, Ramat-Gan (IL).	<b>(74) Agents:</b> BILLINGS, Lucy, J. et al.; Incyte Pharmaceuticals, Inc., 3174 Porter Drive, Palo Alto, CA 94304 (US). <b>(81) Designated States:</b> AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, HR, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG). <b>Published</b> <i>Without international search report and to be republished upon receipt of that report.</i>	
<b>(54) Title:</b> PROSTATE CANCER-ASSOCIATED GENES  <b>(57) Abstract</b>  The invention provides novel prostate cancer-associated genes and polypeptides encoded by those genes. The invention also provides expression vectors, host cells, antibodies, agonists, and antagonists. The invention also provides methods for diagnosing, treating or preventing diseases.		

**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

## PROSTATE CANCER-ASSOCIATED GENES

A portion of the disclosure of this patent document contains material which is  
5 subject to copyright protection. The copyright owner has no objection to the facsimile  
reproduction by anyone of the patent document or the patent disclosure, as it appears in the  
Patent and Trademark Office patent file or records, but otherwise reserves all copyright  
rights whatsoever.

10

## TECHNICAL FIELD

The invention relates to a method for analyzing gene expression patterns. The  
invention also relates to eight prostate cancer-associated genes identified by the method  
and their corresponding polypeptides and to the use of these biomolecules in diagnosis,  
15 prognosis, treatment, prevention, and evaluation of therapies for diseases, particularly  
diseases associated with cell proliferation, such as cancer.

## BACKGROUND OF THE INVENTION

20 The DNA sequences of many human genes have been determined, but for many of  
these genes, their biological function, and in particular their relationship to disease, is  
unknown or poorly understood. Current laboratory and computational methods to  
determine or predict the possible functions of newly-sequenced genes are slow and  
expensive. Thus, new methods that provide additional information on function are  
25 desirable.

Prostate cancer is a common malignancy in men over the age of 50, and the  
incidence increases with age. In the US, there are approximately 132,000 newly  
diagnosed cases of prostate cancer and more than 33,000 deaths from prostate cancer each  
year. The occurrences of prostate cancer vary among different regions in the world. For  
30 example, there are 14 deaths per 100,000 men per year in the US, compared with 22 in  
Sweden and 2 in Japan.

Genes known to be involved in prostate cancer, such as prostate-specific antigen

(PSA), prostatic acid phosphatase (PAP), kallikrein, seminal plasma protein, and prostate-specific transglutaminase, have been used or proposed as the basis for diagnostic and prognostic tests as well as therapeutic targets. In particular, prostate-specific antigen (PSA) is a protease used in diagnosis for prostate cancer (Morris, D. L. et al. (1998) J. Clin. Lab. Anal. 12: 65-74). Prostatic acid phosphatase (PAP) is a phosphomonoesterase synthesized in the prostate and secreted into the seminal plasma under androgenic control (Ostrowski, W. S. and R. Kuciel (1994) Clin. Chim. Acta 226:121-129), and has been used in diagnostic tests for prostate cancer and in prognostic tests for metastatic cancer (Presti, J. C., Jr. and P. R. Carroll (1996) Semin Urol Oncol 14(3): 134-138). Kallikrein is a protease expressed specifically in the prostate and has 80% sequence similarity with PSA (Corey, E., K. R. et al. (1997) Urology 50: 567-572). Kallikrein is being evaluated for use in diagnostic tests for prostate cancer (Pannek, J. and Partin, A. W. (1997) Oncology 11: 1273-1282). Seminal plasma protein is a prostate-specific secreted protein with activity similar to inhibin, a member of the transforming growth factor superfamily implicated in prostate cancer (Mbikay, M., S. et al. (1987) DNA 6: 23-29; Thomas, T. Z. et al. (1998) Prostate 34: 34-43); deletion of the inhibin alpha gene in male rats results in development of primary gonadal granulosa/Sertoli cell tumors (Mellor, S. L. et al. (1998) J. Clin. Endocrinol. Metab. 83: 969-975). Prostate-specific transglutaminase catalyzes post-translational protein cross-linking, and exhibits differential expression in prostate cancer cell lines (Dubbink, H. J. (1996) Biochem. J. 315: 901-908).

The diagnostic sensitivity and specificity and the prognostic accuracy of the tests based on the known genes are substantially less than 100 percent. For example, about 20 percent of the patients undergoing prostatectomy for prostate cancer have normal levels of PSA (Presti and Carroll, supra). Therefore, identification of novel genes and polypeptides that are markers of and potential therapeutic targets for prostate cancer is desirable.

The present invention satisfies a need in the art by providing new compositions which are useful in diagnosis, prognosis, treatment, prevention, and evaluation of therapies for diseases, particularly diseases associated with cell proliferation, such as cancer. We have implemented a method for analyzing gene expression patterns and have identified eight human prostate cancer-associated genes by their coexpression with known prostate cancer-specific genes.

## SUMMARY OF THE INVENTION

In one aspect, the present invention provides a method for identifying biomolecules, such as polynucleotides or polypeptides, useful in the diagnosis, prognosis, 5 treatment, prevention, and evaluation of therapies for diseases, particularly diseases associated with cell proliferation such as cancer, more particularly prostate cancer. The method can also be employed for elucidating genes involved in a common regulatory pathway.

The method comprises first characterizing expression patterns of polynucleotides 10 that are expressed in a plurality of cDNA libraries. The expressed polynucleotides comprise genes of known and unknown functions. Second, the expression patterns of one or more function-specific genes are compared with the expression patterns of one or more of the genes of unknown function to identify a subset of novel genes which have similar expression patterns to those of the function-specific genes.

15 The method compares the expression pattern of two genes by first generating an occurrence vector for each gene. The vector comprises entries for each gene wherein a gene's presence in a cDNA library is represented by a one and a gene's absence by a zero. The vectors are then analyzed to determine whether the expression patterns of any of the genes are similar. Expression patterns are similar if a particular coexpression probability 20 threshold is met. Preferably, the coexpression probability threshold is less than 0.001, and more preferably less than 0.00001.

In a preferred embodiment, the function-specific genes are prostate cancer-specific gene sequences including prostate-specific antigen (PSA), prostatic acid phosphatase (PAP), kallikrein, seminal plasma protein, prostate-specific transglutaminase, and the like. 25 These prostate cancer-specific genes are used to identify other polynucleotides of unidentified function that are predominantly coexpressed with the prostate cancer-specific genes. The polynucleotides analyzed by the present invention can be expressed sequence tags (ESTs), assembled sequences, full length gene coding sequences, introns, regulatory regions, 5' untranslated regions, 3' untranslated regions and the like.

30 In a second aspect, the invention entails a substantially purified polynucleotide identified by the method of the present invention as being associated with prostate cancer. In particular, the polynucleotide comprises a sequence selected from the group consisting

of SEQ ID NOs: 1-8 or its complement or a variant having at least 70% sequence identity to SEQ ID NOs: 1-8 or a polynucleotide that hybridizes under stringent conditions to SEQ ID NOs: 1-8 or a polynucleotide encoding SEQ ID NOs: 9 and 10. The present invention also entails a polynucleotide comprising at least 18 consecutive nucleotides of a sequence  
5 provided above. The polynucleotide is suitable for use in diagnosis, treatment, prognosis, or prevention of a cancer, and in particular, prostate cancer. The polynucleotide is also suitable for the evaluation of therapies for cancer.

In another aspect, the invention provides an expression vector comprising a polynucleotide described above, a host cell comprising the expression vector, and a  
10 method for detecting a target polynucleotide in a sample.

In a further aspect, the invention provides a substantially purified polypeptide comprising an amino acid sequence selected from the group consisting of SEQ ID NO:9 and SEQ ID NO:10. The invention also provides a substantially purified polypeptide having at least 85% identity to SEQ ID NOs:9-10. Additionally, the invention also  
15 provides a sequence with at least 6 sequential amino acids of SEQ ID NOs:9-10.

The invention also provides a method for producing a substantially purified polypeptide comprising the amino acid sequence referred to above, and antibodies, agonists, and antagonists which specifically bind to the polypeptide. Pharmaceutical compositions comprising the polynucleotides or polypeptides of the invention are also  
20 contemplated. Methods for producing a polypeptide of the invention and methods for detecting a target polynucleotide complementary to a polynucleotide of the invention are also included.

In a general aspect, the invention entails a method for identifying biomolecules useful in the diagnosis or treatment of a disease or condition. The method comprises a)  
25 examining expression patterns of a plurality of biomolecules that are expressed in a plurality of cDNA libraries, said expressed biomolecules comprising one or more disease-specific biomolecules and one or more biomolecules of unknown function; and b)  
comparing the expression patterns of said disease-specific biomolecules with the expression patterns of the biomolecules of unknown function to identify a subset of the  
30 biomolecules of unknown function which have similar expression patterns to those of disease-specific biomolecules.

## BRIEF DESCRIPTION OF THE SEQUENCE LISTING

The Sequence Listing provides exemplary prostate cancer-associated sequences including polynucleotide sequences, SEQ ID NOs: 1-8, and polypeptide sequences, SEQ ID NOs: 9-10. Each sequence is identified by a sequence identification number (SEQ ID NO) and by the Incyte Clone number from which the sequence was first identified.

## DESCRIPTION OF THE INVENTION

It must be noted that as used herein and in the appended claims, the singular forms “a,” “an,” and “the” include the plural reference unless the context clearly dictates otherwise. Thus, for example, a reference to “a host cell” includes a plurality of such host cells, and a reference to “an antibody” is a reference to one or more antibodies and equivalents thereof known to those skilled in the art, and so forth.

## DEFINITIONS

“NSEQ” refers generally to a polynucleotide sequence of the present invention, including SEQ ID NOs: 1-8. “PSEQ” refers generally to a polypeptide sequence of the present invention, including SEQ ID NOs: 9-10.

A “variant” refers to either a polynucleotide or a polypeptide whose sequence diverges from SEQ ID NOs: 1-10 or SEQ ID NOs: 9-10, respectively. Polynucleotide sequence divergence may result from mutational changes such as deletions, additions, and substitutions of one or more nucleotides; it may also occur because of differences in codon usage. Each of these types of changes may occur alone, or in combination, one or more times in a given sequence. Polypeptide variants include sequences that possess at least one structural or functional characteristic of SEQ ID NOs: 9-10.

“Gene” or “gene sequence” refers to the partial or complete coding sequence of a gene. The term also refers to 5' or 3' untranslated regions. The gene may be in a sense or antisense (complementary) orientation.

“Prostate cancer-specific gene” refers to a gene sequence which has been previously identified as useful in the diagnosis, treatment, prognosis, or prevention of prostate cancer. Typically, this means that the prostate cancer-specific gene is expressed at higher levels in prostate cancer tissue when compared with healthy tissue.

"Prostate cancer-associated gene" refers to a gene sequence whose expression pattern is similar to that of the prostate cancer-specific genes and which are useful in the diagnosis, treatment, prognosis, or prevention of cancer. The gene sequences can also be used in the evaluation of therapies for cancer.

- 5 "Substantially purified" refers to a nucleic acid or an amino acid sequence that is removed from its natural environment and is isolated or separated, and is at least about 60% free, preferably about 75% free, and most preferably about 90% free from other components with which it is naturally present.

## 10 THE INVENTION

- The present invention encompasses a method for identifying biomolecules that are associated with a specific disease, regulatory pathway, subcellular compartment, cell type, tissue type, or species. In particular, the method identifies gene sequences useful in diagnosis, prognosis, treatment, prevention, and evaluation of therapies for diseases  
15 associated with cell proliferation, particularly cancer, and more particularly prostate cancer.

- The method entails first identifying polynucleotides that are expressed in the cDNA libraries. The polynucleotides include genes of known function, genes known to be specifically expressed in a specific disease process, subcellular compartment, cell type,  
20 tissue type, or species. Additionally, the polynucleotides include genes of unknown function. The expression patterns of the known genes are then compared with those of the genes of unknown function to determine whether a specified coexpression probability threshold is met. Through this comparison, a subset of the polynucleotides having a high coexpression probability with the known genes can be identified. The high coexpression  
25 probability correlates with a particular coexpression probability threshold which is less than 0.001, and more preferably less than 0.00001.

- The polynucleotides originate from cDNA libraries derived from a variety of sources including, but not limited to, eukaryotes such as human, mouse, rat, dog, monkey, plant, and yeast and prokaryotes such as bacteria and viruses. These polynucleotides can  
30 also be selected from a variety of sequence types including, but not limited to, expressed sequence tags (ESTs), assembled polynucleotide sequences, full length gene coding regions, introns, regulatory sequences, 5' untranslated regions, and 3' untranslated regions.



To have statistically significant analytical results, the polynucleotides need to be expressed in at least three cDNA libraries.

The cDNA libraries used in the coexpression analysis of the present invention can be obtained from blood vessels, heart, blood cells, cultured cells, connective tissue, 5 epithelium, islets of Langerhans, neurons, phagocytes, biliary tract, esophagus, gastrointestinal system, liver, pancreas, fetus, placenta, chromaffin system, endocrine glands, ovary, uterus, penis, prostate, seminal vesicles, testis, bone marrow, immune system, cartilage, muscles, skeleton, central nervous system, ganglia, neuroglia, neurosecretory system, peripheral nervous system, bronchus, larynx, lung, nose, pleurus, 10 ear, eye, mouth, pharynx, exocrine glands, bladder, kidney, ureter, and the like. The number of cDNA libraries selected can range from as few as 20 to greater than 10,000. Preferably, the number of the cDNA libraries is greater than 500.

In a preferred embodiment, gene sequences are assembled to reflect related sequences, such as assembled sequence fragments derived from a single transcript. 15 Assembly of the polynucleotide sequences can be performed using sequences of various types including, but not limited to, ESTs, extensions, or shotgun sequences. In a most preferred embodiment, the polynucleotide sequences are derived from human sequences that have been assembled using the algorithm disclosed in "Database and System for Storing, Comparing and Displaying Related Biomolecular Sequence Information", 20 Lincoln et al., Serial No:60/079,469, filed March 26, 1998, herein incorporated by reference.

Experimentally, differential expression of the polynucleotides can be evaluated by methods including, but not limited to, differential display by spatial immobilization or by gel electrophoresis, genome mismatch scanning, representational difference analysis, and 25 transcript imaging. Additionally, differential expression can be assessed by microarray technology. These methods may be used alone or in combination.

Genes known to be prostate cancer-specific can be selected based on the use of the genes as diagnostic or prognostic markers or as therapeutic targets for prostate cancer. Preferably, the prostate cancer-specific genes include prostate-specific antigen (PSA), 30 prostatic acid phosphatase (PAP), kallikrein, seminal plasma protein, prostate-specific transglutaminase, and the like.

The procedure for identifying novel genes that exhibit a statistically significant

coexpression pattern with prostate cancer-specific genes is as follows. First, the presence or absence of a gene sequence in a cDNA library is defined: a gene is present in a cDNA library when at least one cDNA fragment corresponding to that gene is detected in a cDNA sample taken from the library, and a gene is absent from a library when no

5 corresponding cDNA fragment is detected in the sample.

Second, the significance of gene coexpression is evaluated using a probability method to measure a due-to-chance probability of the coexpression. The probability method can be the Fisher exact test, the chi-squared test, or the kappa test. These tests and examples of their applications are well known in the art and can be found in standard  
 10 statistics texts (Agresti, A. (1990) Categorical Data Analysis. New York, NY, Wiley; Rice, J. A. (1988) Mathematical Statistics and Data Analysis. Pacific Grove, CA, Wadsworth & Brooks/Cole). A Bonferroni correction (Rice, supra, page 384) can also be applied in combination with one of the probability methods for correcting statistical results of one gene versus multiple other genes. In a preferred embodiment, the due-to-chance  
 15 probability is measured by a Fisher exact test, and the threshold of the due-to-chance probability is set to less than 0.001, more preferably less than 0.00001.

To determine whether two genes, A and B, have similar coexpression patterns, occurrence data vectors can be generated as illustrated in Table 1, wherein a gene's presence is indicated by a one and its absence by a zero. A zero indicates that the gene  
 20 did not occur in the library, and a one indicates that it occurred at least once.

Table 1. Occurrence data for genes A and B

	Library 1	Library 2	Library 3	...	Library N
25 gene A	1	1	0	...	0
gene B	1	0	1	...	0

For a given pair of genes, the occurrence data in Table 1 can be summarized in a 2x2  
 30 contingency table.

Table 2. Contingency table for co-occurrences of genes A and B

	Gene A present	Gene A absent	Total
Gene B present	8	2	10
Gene B absent	2	18	20
5 Total	10	20	30

Table 2 presents co-occurrence data for gene A and gene B in a total of 30 libraries. Both gene A and gene B occur 10 times in the libraries. Table 2 summarizes and  
 10 presents 1) the number of times gene A and B are both present in a library, 2) the number of times gene A and B are both absent in a library, 3) the number of times gene A is present while gene B is absent, and 4) the number of times gene B is present while gene A is absent. The upper left entry is the number of times the two genes co-occur in a library, and the middle right entry is the number of times neither gene occurs in a library. The off  
 15 diagonal entries are the number of times one gene occurs while the other does not. Both A and B are present eight times and absent 18 times, gene A is present while gene B is absent two times, and gene B is present while gene A is absent two times. The probability ("p-value") that the above association occurs due to chance as calculated using a Fisher exact test is 0.0003. Associations are generally considered significant if a p-value is less  
 20 than 0.01 (Agresti, *supra*; Rice, *supra*).

This method of estimating the probability for coexpression of two genes makes several assumptions. The method assumes that the libraries are independent and are identically sampled. However, in practical situations, the selected cDNA libraries are not entirely independent because more than one library may be obtained from a single patient  
 25 or tissue, and they are not entirely identically sampled because different numbers of cDNA's may be sequenced from each library (typically ranging from 5,000 to 10,000 cDNA's per library). In addition, because a Fisher exact coexpression probability is calculated for each gene versus 41,419 other genes, a Bonferroni correction for multiple statistical tests is necessary.

30 Using the method of the present invention, we have identified eight novel genes that exhibit strong association, or coexpression, with known genes that are prostate cancer-

specific. These prostate cancer-specific genes include glandular kallikrein, prostate seminal protein, prostate-specific antigen, and prostatic acid phosphatase. The results presented in Tables 5 to 12 show that the expression of eight novel genes have direct or indirect association with the expression of cancer-specific genes, in particular prostate cancer-specific genes. Therefore, the novel genes can potentially be used in diagnosis, treatment, prognosis, or prevention of cancer, or in the evaluation of therapies for cancer. Further, the gene products of the eight novel genes are potential therapeutic proteins and targets of anti-cancer therapeutics.

Therefore, in one embodiment, the present invention encompasses a polynucleotide sequence comprising the sequence of SEQ ID NOs:1-8. These eight polynucleotides are shown by the method of the present invention to have strong coexpression association with prostate cancer-specific genes and with each other. The invention also encompasses a variant of the polynucleotide sequence, its complement, or 18 consecutive nucleotides of the sequences provided in the above described sequences. Variant polynucleotide sequences typically have at least about 70%, more preferably at least about 85%, and most preferably at least about 95% polynucleotide sequence identity to NSEQ.

One preferred method for identifying variants entails using NSEQ and/or PSEQ sequences to search against the GenBank primate (pri), rodent (rod), and mammalian (mam), vertebrate (vrtp), and eukaryote (eukp) databases, SwissProt, BLOCKS (Bairoch, A. et al. (1997) *Nucleic Acids Res.* 25:217-221), PFAM, and other databases that contain previously identified and annotated motifs, sequences, and gene functions. Methods that search for primary sequence patterns with secondary structure gap penalties (Smith, T. et al. (1992) *Protein Engineering* 5:35-51) as well as algorithms such as BLAST (Basic Local Alignment Search Tool; Altschul, S.F. (1993) *J. Mol. Evol* 36:290-300; and Altschul et al. (1990) *J. Mol. Biol.* 215:403-410), BLOCKS (Henikoff S. and Henikoff G.J. (1991) *Nucleic Acids Research* 19:6565-6572), Hidden Markov Models (HMM; Eddy, S.R. (1996) *Cur. Opin. Str. Biol.* 6:361-365; and Sonnhammer, E.L.L. et al. (1997) *Proteins* 28:405-420), and the like, can be used to manipulate and analyze nucleotide and amino acid sequences. These databases, algorithms and other methods are well known in the art and are described in Ausubel, F.M. et al. (1997; Short Protocols in Molecular Biology, John Wiley & Sons, New York, NY) and in Meyers, R.A. (1995; Molecular Biology and Biotechnology, Wiley VCH, Inc, New York, NY, p 856-853).

Also encompassed by the invention are polynucleotide sequences that are capable of hybridizing to SEQ ID NOs:1-8, and fragments thereof under stringent conditions.

Stringent conditions can be defined by salt concentration, temperature, and other chemicals and conditions well known in the art. In particular, stringency can be increased

5 by reducing the concentration of salt, or raising the hybridization temperature.

For example, stringent salt concentration will ordinarily be less than about 750 mM NaCl and 75 mM trisodium citrate, preferably less than about 500 mM NaCl and 50 mM trisodium citrate, and most preferably less than about 250 mM NaCl and 25 mM trisodium citrate. Stringent temperature conditions will ordinarily include temperatures of at least  
10 about 30°C, more preferably of at least about 37°C, and most preferably of at least about 42°C. Varying additional parameters, such as hybridization time, the concentration of detergent (sodium dodecyl sulfate, SDS) or solvent (formamide), and the inclusion or exclusion of carrier DNA, are well known to those skilled in the art. Additional variations on these conditions will be readily apparent to those skilled in the art (Wahl, G.M. and  
15 S.L. Berger (1987) *Methods Enzymol.* 152:399-407; Kimmel, A.R. (1987) *Methods Enzymol.* 152:507-511; Ausubel, F.M. et al. (1997) Short Protocols in Molecular Biology, John Wiley & Sons, New York, NY; and Sambrook, J. et al. (1989) Molecular Cloning. A Laboratory Manual, Cold Spring Harbor Press, Plainview, NY).

NSEQ or the polynucleotide sequences encoding PSEQ can be extended utilizing a  
20 partial nucleotide sequence and employing various PCR-based methods known in the art to detect upstream sequences, such as promoters and regulatory elements. (See, e.g., Dieffenbach, C.W. and G.S. Dveksler (1995; PCR Primer, a Laboratory Manual, Cold Spring Harbor Press, Plainview, NY, pp.1-5; Sarkar, G. (1993; *PCR Methods Applic.* 2:318-322); Triglia, T. et al. (1988; *Nucleic Acids Res.* 16:8186); Lagerstrom, M. et al.  
25 (1991; *PCR Methods Applic.* 1:111-119); and Parker, J.D. et al. (1991; *Nucleic Acids Res.* 19:3055-306). Additionally, one may use PCR, nested primers, and PROMOTERFINDER libraries to walk genomic DNA (Clontech, Palo Alto, CA). This procedure avoids the need to screen libraries and is useful in finding intron/exon junctions. For all PCR-based methods, primers may be designed using commercially available  
30 software, such as OLIGO 4.06 Primer Analysis software (National Biosciences Inc., Plymouth MN) or another appropriate program, to be about 18 to 30 nucleotides in length, to have a GC content of about 50% or more, and to anneal to the template at temperatures

of about 68°C to 72°C.

In another aspect of the invention, NSEQ or the polynucleotide sequences encoding PSEQ can be cloned in recombinant DNA molecules that direct expression of PSEQ or the polypeptides encoded by NSEQ, or structural or functional fragments thereof, in appropriate host cells. Due to the inherent degeneracy of the genetic code, other DNA sequences which encode substantially the same or a functionally equivalent amino acid sequence may be produced and used to express the polypeptides of PSEQ or the polypeptides encoded by NSEQ. The nucleotide sequences of the present invention can be engineered using methods generally known in the art in order to alter the nucleotide sequences for a variety of purposes including, but not limited to, modification of the cloning, processing, and/or expression of the gene product. DNA shuffling by random fragmentation and PCR reassembly of gene fragments and synthetic oligonucleotides may be used to engineer the nucleotide sequences. For example, oligonucleotide-mediated site-directed mutagenesis may be used to introduce mutations that create new restriction sites, alter glycosylation patterns, change codon preference, produce splice variants, and so forth.

In order to express a biologically active polypeptide encoded by NSEQ, NSEQ or the polynucleotide sequences encoding PSEQ, or derivatives thereof, may be inserted into an appropriate expression vector, i.e., a vector which contains the necessary elements for transcriptional and translational control of the inserted coding sequence in a suitable host. These elements include regulatory sequences, such as enhancers, constitutive and inducible promoters, and 5' and 3' untranslated regions in the vector and in NSEQ or polynucleotide sequences encoding PSEQ. Methods which are well known to those skilled in the art may be used to construct expression vectors containing NSEQ or polynucleotide sequences encoding PSEQ and appropriate transcriptional and translational control elements. These methods include in vitro recombinant DNA techniques, synthetic techniques, and in vivo genetic recombination. (See, e.g., Sambrook (supra) and Ausubel, (supra).

A variety of expression vector/host cell systems may be utilized to contain and express NSEQ or polynucleotide sequences encoding PSEQ. These include, but are not limited to, microorganisms such as bacteria transformed with recombinant bacteriophage, plasmid, or cosmid DNA expression vectors; yeast transformed with yeast expression

vectors; insect cell systems infected with viral expression vectors (baculovirus); plant cell systems transformed with viral expression vectors, cauliflower mosaic virus (CaMV) or tobacco mosaic virus (TMV), or with bacterial expression vectors (Ti or pBR322 plasmids); or animal cell systems. The invention is not limited by the host cell employed.

- 5 For long term production of recombinant proteins in mammalian systems, stable expression of a polypeptide encoded by NSEQ in cell lines is preferred. For example, NSEQ or sequences encoding PSEQ can be transformed into cell lines using expression vectors which may contain viral origins of replication and/or endogenous expression elements and a selectable marker gene on the same or on a separate vector.

- 10 In general, host cells that contain NSEQ and that express PSEQ may be identified by a variety of procedures known to those of skill in the art. These procedures include, but are not limited to, DNA-DNA or DNA-RNA hybridizations, PCR amplification, and protein bioassay or immunoassay techniques which include membrane, solution, or chip based technologies for the detection and/or quantification of nucleic acid or protein
- 15 sequences. Immunological methods for detecting and measuring the expression of PSEQ using either specific polyclonal or monoclonal antibodies are known in the art. Examples of such techniques include enzyme-linked immunosorbent assays (ELISAs), radioimmunoassays (RIAs), and fluorescence activated cell sorting (FACS).

- Host cells transformed with NSEQ or polynucleotide sequences encoding PSEQ
- 20 may be cultured under conditions suitable for the expression and recovery of the protein from cell culture. The protein produced by a transformed cell may be secreted or retained intracellularly depending on the sequence and/or the vector used. As will be understood by those of skill in the art, expression vectors containing polynucleotides of NSEQ or polynucleotides encoding PSEQ may be designed to contain signal sequences which direct
- 25 secretion of PSEQ or polypeptides encoded by NSEQ through a prokaryotic or eukaryotic cell membrane.

- In addition, a host cell strain may be chosen for its ability to modulate expression of the inserted sequences or to process the expressed protein in the desired fashion. Such modifications of the polypeptide include, but are not limited to, acetylation, carboxylation,
- 30 glycosylation, phosphorylation, lipidation, and acylation. Post-translational processing which cleaves a "prepro" form of the protein may also be used to specify protein targeting, folding, and/or activity. Different host cells which have specific cellular machinery and

characteristic mechanisms for post-translational activities (e.g., CHO, HeLa, MDCK, HEK293, and WI38), are available from the American Type Culture Collection (ATCC, Bethesda, MD) and may be chosen to ensure the correct modification and processing of the foreign protein.

5 In another embodiment of the invention, natural, modified, or recombinant NSEQ or nucleic acid sequences encoding PSEQ are ligated to a heterologous sequence resulting in translation of a fusion protein containing heterologous protein moieties in any of the aforementioned host systems. Such heterologous protein moieties facilitate purification of fusion proteins using commercially available affinity matrices. Such moieties include, but are not limited to, glutathione S-transferase (GST), maltose binding protein (MBP),  
10 thioredoxin (Trx), calmodulin binding peptide (CBP), 6-His, FLAG, *c-myc*, hemagglutinin (HA) and monoclonal antibody epitopes..

In another embodiment, NSEQ or sequences encoding PSEQ are synthesized, in whole or in part, using chemical methods well known in the art. (See, e.g., Caruthers,  
15 M.H. et al. (1980) Nucl. Acids Res. Symp. Ser. 215-223; Horn, T. et al. (1980) Nucl. Acids Res. Symp. Ser. 225-232; and Ausubel, supra). Alternatively, PSEQ or a polypeptide sequence encoded by NSEQ itself, or a fragment thereof, may be synthesized using chemical methods. For example, peptide synthesis can be performed using various solid-phase techniques (Roberge, J.Y. et al. (1995) Science 269:202-204). Automated  
20 synthesis may be achieved using the ABI 431A Peptide Synthesizer (Perkin Elmer). Additionally, PSEQ or the amino acid sequence encoded by NSEQ, or any part thereof, may be altered during direct synthesis and/or combined with sequences from other proteins, or any part thereof, to produce a polypeptide variant.

In another embodiment, the invention entails a substantially purified polypeptide  
25 comprising the amino acid sequence selected from the group consisting of SEQ ID NO:9, SEQ ID NO:10, or fragments thereof. SEQ ID NO:9 is encoded by SEQ ID NO:4 and is a potential transmembrane protein which interacts with a cell surface receptor. SEQ ID NO:10 is encoded by SEQ ID NO:8 and has potential sequence homology with a family of GPI-linked cell-surface glycoproteins, Ly-6/u-PAR.

30

#### **DIAGNOSTICS and THERAPEUTICS**

The sequences of these genes can be used in diagnosis, prognosis, treatment,



prevention, and evaluation of therapies for diseases associated with cell proliferation, particularly cancer, and more particularly prostate cancer. Further, the amino acid sequences encoded by the novel genes are potential therapeutic proteins and targets of anti-cancer therapeutics.

5 In one preferred embodiment, the polynucleotide sequences of NSEQ or the polynucleotides encoding PSEQ are used for diagnostic purposes to determine the absence, presence, and excess expression of PSEQ, and to monitor regulation of the levels of mRNA or the polypeptides encoded by NSEQ during therapeutic intervention. The polynucleotides may be at least 18 nucleotides long, complementary RNA and DNA  
10 molecules, branched nucleic acids, and peptide nucleic acids (PNAs). Alternatively, the polynucleotides are used to detect and quantitate gene expression in samples in which expression of PSEQ or the polypeptides encoded by NSEQ are correlated with disease. Additionally, NSEQ or the polynucleotides encoding PSEQ can be used to detect genetic polymorphisms associated with a disease. These polymorphisms may be detected at the  
15 transcript cDNA or genomic level.

The specificity of the probe, whether it is made from a highly specific region, e.g., the 5' regulatory region, or from a less specific region, e.g., a conserved motif, and the stringency of the hybridization or amplification (maximal, high, intermediate, or low), will determine whether the probe identifies only naturally occurring sequences encoding  
20 PSEQ, allelic variants, or related sequences.

Probes may also be used for the detection of related sequences, and should preferably have at least 50% sequence identity to any of the NSEQ or PSEQ-encoding sequences.

Means for producing specific hybridization probes for DNAs encoding PSEQ  
25 include the cloning of NSEQ or polynucleotide sequences encoding PSEQ into vectors for the production of mRNA probes. Such vectors are known in the art, are commercially available, and may be used to synthesize RNA probes in vitro by means of the addition of the appropriate RNA polymerases and the appropriate labeled nucleotides. Hybridization probes may be labeled by a variety of reporter groups, for example, by radionuclides such  
30 as  $^{32}\text{P}$  or  $^{35}\text{S}$ , or by enzymatic labels, such as alkaline phosphatase coupled to the probe via avidin/biotin coupling systems, by fluorescent labels and the like. The polynucleotide sequences encoding PSEQ may be used in Southern or northern analysis, dot blot, or other

membrane-based technologies; in PCR technologies; and in microarrays utilizing fluids or tissues from patients to detect altered PSEQ expression. Such qualitative or quantitative methods are well known in the art.

NSEQ or the nucleotide sequences encoding PSEQ can be labeled by standard  
5 methods and added to a fluid or tissue sample from a patient under conditions suitable for the formation of hybridization complexes. After a suitable incubation period, the sample is washed and the signal is quantitated and compared with a standard value. If the amount of signal in the patient sample is significantly altered in comparison to the standard value then the presence of altered levels of nucleotide sequences of NSEQ and those encoding  
10 PSEQ in the sample indicates the presence of the associated disease. Such assays may also be used to evaluate the efficacy of a particular therapeutic treatment regimen in animal studies, in clinical trials, or to monitor the treatment of an individual patient.

Once the presence of a disease is established and a treatment protocol is initiated, hybridization or amplification assays can be repeated on a regular basis to determine if the  
15 level of expression in the patient begins to approximate that which is observed in the normal subject. The results obtained from successive assays may be used to show the efficacy of treatment over a period ranging from several days to months.

The polynucleotides may be used for the diagnosis of a variety of diseases associated with cell proliferation including cancer such as adenocarcinoma, leukemia,  
20 lymphoma, melanoma, myeloma, sarcoma, teratocarcinoma, and, in particular, cancers of the adrenal gland, bladder, bone, bone marrow, brain, breast, cervix, gall bladder, ganglia, gastrointestinal tract, heart, kidney, liver, lung, muscle, ovary, pancreas, parathyroid, penis, prostate, salivary glands, skin, spleen, testis, thymus, thyroid, and uterus.

Alternatively, the polynucleotides may be used as targets in a microarray. The  
25 microarray can be used to monitor the expression level of large numbers of genes simultaneously and to identify splice variants, mutations, and polymorphisms. This information may be used to determine gene function, to understand the genetic basis of a disease, to diagnose a disease, and to develop and monitor the activities of therapeutic agents.

30 In yet another alternative, polynucleotides may be used to generate hybridization probes useful in mapping the naturally occurring genomic sequence. Fluorescent in situ hybridization (FISH) may be correlated with other physical chromosome mapping

techniques and genetic map data. (See, e.g., Heinz-Ulrich, et al. (1995) in Meyers, R.A. (ed.) Molecular Biology and Biotechnology, VCH Publishers New York, NY, pp. 965-968).

In another embodiment, antibodies which specifically bind PSEQ may be used for  
5 the diagnosis of diseases characterized by the over-or-underexpression of PSEQ or  
polypeptides encoded by NSEQ. Alternatively, one may use competitive drug screening  
assays in which neutralizing antibodies capable of binding PSEQ or the polypeptides  
encoded by NSEQ specifically compete with a test compound for binding the  
polypeptides. In this manner, antibodies can be used to detect the presence of any peptide  
10 which shares one or more antigenic determinants with PSEQ or the polypeptides encoded  
by NSEQ. Diagnostic assays for PSEQ or the polypeptides encoded by NSEQ include  
methods which utilize the antibody and a label to detect PSEQ or the polypeptided  
encoded by NSEQ in human body fluids or in extracts of cells or tissues. A variety of  
protocols for measuring PSEQ or the polypeptides encoded by NSEQ, including ELISAs,  
15 RIAs, and FACS, are well known in the art and provide a basis for diagnosing altered or  
abnormal levels of the expression of PSEQ or the polypeptides encoded by NSEQ.  
Normal or standard values for PSEQ expression are established by combining body fluids  
or cell extracts taken from normal subjects, preferably human, with antibody to PSEQ or a  
polypeptide encoded by NSEQ under conditions suitable for complex formation. The  
20 amount of standard complex formation may be quantitated by various methods, preferably  
by photometric means. Quantities of PSEQ or the polypeptides encoded by NSEQ  
expressed in subject, control, and disease samples from biopsied tissues are compared with  
the standard values. Deviation between standard and subject values establishes the  
parameters for diagnosing or monitoring disease.

25 In another aspect, the polynucleotides and polypeptides of the present invention  
can be employed for treatment or the monitoring of therapeutic treatments for cancers.  
The polynucleotides of NSEQ or those encoding PSEQ, or any fragment or complement  
thereof, may be used for therapeutic purposes. In one aspect, the complement of the  
polynucleotides of NSEQ or those encoding PSEQ may be used in situations in which it  
30 would be desirable to block the transcription or translation of the mRNA.

Expression vectors derived from retroviruses, adenoviruses, or herpes or vaccinia  
viruses, or from various bacterial plasmids, may be used for delivery of nucleotide

sequences to the targeted organ, tissue, or cell population. Methods which are well known to those skilled in the art can be used to construct vectors to express nucleic acid sequences complementary to the polynucleotides encoding PSEQ. (See, e.g., Sambrook, supra; and Ausubel, supra.)

5 Genes having polynucleotide sequences of NSEQ or those encoding PSEQ can be turned off by transforming a cell or tissue with expression vectors which express high levels of a polynucleotide, or fragment thereof, encoding PSEQ. Such constructs may be used to introduce untranslatable sense or antisense sequences into a cell. Oligonucleotides derived from the transcription initiation site, e.g., between about positions -10 and +10  
10 from the start site, are preferred. Similarly, inhibition can be achieved using triple helix base-pairing methodology. Triple helix pairing is useful because it causes inhibition of the ability of the double helix to open sufficiently for the binding of polymerases, transcription factors, or regulatory molecules. Recent therapeutic advances using triplex DNA have been described in the literature. (See, e.g., Gee, J.E. et al. (1994) in Huber,  
15 B.E. and B.I. Carr, Molecular and Immunologic Approaches, Futura Publishing Co., Mt. Kisco, NY, pp. 163-177.) Ribozymes, enzymatic RNA molecules, may also be used to catalyze the specific cleavage of RNA.

RNA molecules may be modified to increase intracellular stability and half-life. Possible modifications include, but are not limited to, the addition of flanking sequences at  
20 the 5' and/or 3' ends of the molecule, or the use of phosphorothioate or 2' O-methyl rather than phosphodiesterase linkages within the backbone of the molecule. This concept is inherent in the production of PNAs and can be extended in all of these molecules by the inclusion of nontraditional bases such as inosine, queosine, and wybutosine, as well as acetyl-, methyl-, thio-, and similarly modified forms of adenine, cytidine, guanine,  
25 thymine, and uridine which are not as easily recognized by endogenous endonucleases.

Many methods for introducing vectors into cells or tissues are available and equally suitable for use in vivo, in vitro, and ex vivo. For ex vivo therapy, vectors may be introduced into stem cells taken from the patient and clonally propagated for autologous transplant back into that same patient. Delivery by transfection, by liposome injections, or  
30 by polycationic amino polymers may be achieved using methods which are well known in the art. (See, e.g., Goldman, C.K. et al. (1997) *Nature Biotechnology* 15:462-466.)

Further, an antagonist or antibody of a polypeptide of PSEQ or encoded by NSEQ

may be administered to a subject to treat or prevent a cancer associated with increased expression or activity of PSEQ. An antibody which specifically binds the polypeptide may be used directly as an antagonist or indirectly as a targeting or delivery mechanism for bringing a pharmaceutical agent to cells or tissue which express the the polypeptide.

5       Antibodies to PSEQ or polypeptides encoded by NSEQ may also be generated using methods that are well known in the art. Such antibodies may include, but are not limited to, polyclonal, monoclonal, chimeric, and single chain antibodies, Fab fragments, and fragments produced by a Fab expression library. Neutralizing antibodies (i.e., those which inhibit dimer formation) are especially preferred for therapeutic use. Monoclonal  
10       antibodies to PSEQ may be prepared using any technique which provides for the production of antibody molecules by continuous cell lines in culture. These include, but are not limited to, the hybridoma technique, the human B-cell hybridoma technique, and the EBV-hybridoma technique. In addition, techniques developed for the production of chimeric antibodies can be used. (See, for example, Molecular Biology and  
15       Biotechnology, R.A. Myers, ed.,(1995)John Wiley & Sons, Inc., New York, NY). Alternatively, techniques described for the production of single chain antibodies may be employed. Antibody fragments which contain specific binding sites for PSEQ or the polypeptide sequences encoded by NSEQ may also be generated.

      Various immunoassays may be used for screening to identify antibodies having the  
20       desired specificity. Numerous protocols for competitive binding or immunoradiometric assays using either polyclonal or monoclonal antibodies with established specificities are well known in the art.

      Yet further, an agonist of a polypeptide of PSEQ or that encoded by NSEQ may be administered to a subject to treat or prevent a cancer associated with decreased expression  
25       or activity of the polypeptide.

      An additional aspect of the invention relates to the administration of a pharmaceutical or sterile composition, in conjunction with a pharmaceutically acceptable carrier, for any of the therapeutic effects discussed above. Such pharmaceutical compositions may consist of polypeptides of PSEQ or those encoded by NSEQ, antibodies  
30       to the polypeptides, and mimetics, agonists, antagonists, or inhibitors of the polypeptides. The compositions may be administered alone or in combination with at least one other agent, such as a stabilizing compound, which may be administered in any sterile,

biocompatible pharmaceutical carrier including, but not limited to, saline, buffered saline, dextrose, and water. The compositions may be administered to a patient alone, or in combination with other agents, drugs, or hormones.

The pharmaceutical compositions utilized in this invention may be administered by  
5 any number of routes including, but not limited to, oral, intravenous, intramuscular, intra-arterial, intramedullary, intrathecal, intraventricular, transdermal, subcutaneous, intraperitoneal, intranasal, enteral, topical, sublingual, or rectal means.

In addition to the active ingredients, these pharmaceutical compositions may contain suitable pharmaceutically-acceptable carriers comprising excipients and auxiliaries  
10 which facilitate processing of the active compounds into preparations which can be used pharmaceutically. Further details on techniques for formulation and administration may be found in the latest edition of Remington's Pharmaceutical Sciences (Maack Publishing Co., Easton, PA).

For any compound, the therapeutically effective dose can be estimated initially  
15 either in cell culture assays, e.g., of neoplastic cells or in animal models such as mice, rats, rabbits, dogs, or pigs. An animal model may also be used to determine the appropriate concentration range and route of administration. Such information can then be used to determine useful doses and routes for administration in humans.

A therapeutically effective dose refers to that amount of active ingredient, for  
20 example, polypeptides of PSEQ or those encoded by NSEQ, or fragments thereof, antibodies of the polypeptides, and agonists, antagonists or inhibitors of the polypeptides, which ameliorates the symptoms or condition. Therapeutic efficacy and toxicity may be determined by standard pharmaceutical procedures in cell cultures or with experimental animals, such as by calculating the ED<sub>50</sub> (the dose therapeutically effective in 50% of the  
25 population) or LD<sub>50</sub> (the dose lethal to 50% of the population) statistics.

Any of the therapeutic methods described above may be applied to any subject in need of such therapy, including, for example, mammals such as dogs, cats, cows, horses, rabbits, monkeys, and most preferably, humans.

30

## EXAMPLES

It is understood that this invention is not limited to the particular methodology, protocols, and reagents described, as these may vary. It is also understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to limit the scope of the present invention which will be limited only by the  
5 appended claims. The examples below are provide to illustrate the subject invention and are not included for the purpose of limiting the invention.

## **I. PROSTUT05 cDNA Library Construction**

For purposes of example, the preparation of the PROSTUT05 library is described.  
10 The PROSTUT05 cDNA library was constructed using polyA RNA isolated from prostate tumor tissue removed from a 69-year-old Caucasian male during a radical prostatectomy. Pathology indicated adenocarcinoma (Gleason grade 3+4) involving the right side peripherally. The tumor invaded the capsule but did not extend beyond it; perineural invasion was present. Adenofibromatous hyperplasia was also present. The right seminal  
15 vesicle was involved with tumor. The patient presented with elevated prostate specific antigen (PSA). Patient history included partial colectomy, and tobacco use. Family history included congestive heart failure, multiple myeloma, hyperlipidemia, and rheumatoid arthritis.

The frozen tissue was homogenized and lysed using a Brinkmann Homogenizer  
20 Polytron PT-3000 (Brinkmann Instruments, Westbury, NJ) in guanidinium isothiocyanate solution. The lysate was centrifuged over a 5.7 M CsCl cushion using an Beckman SW28 rotor in a Beckman L8-70M Ultracentrifuge (Beckman Instruments) for 18 hours at 25,000 rpm at ambient temperature. The RNA was extracted with acid phenol pH 4.7, precipitated using 0.3 M sodium acetate and 2.5 volumes of ethanol, resuspended in  
25 RNase-free water, and treated with DNase at 37°C. mRNA extraction and precipitation were repeated as before. The mRNA was isolated with the Qiagen Oligotex kit (QIAGEN, Chatsworth, CA) and used to construct the cDNA library.

The mRNA was handled according to the recommended protocols in the SuperScript Plasmid System for cDNA synthesis and plasmid cloning (GIBCO/BRL).  
30 The cDNAs were fractionated on a Sepharose CL4B column (Pharmacia), and those cDNAs exceeding 400 bp were ligated into pSport I. The plasmid pSport I was subsequently transformed into DH5 $\alpha$ <sup>TM</sup> competent cells (LifeTechnologies, Gaithersburg,

MD).

## II. Isolation and Sequencing of cDNA Clones

Plasmid DNA was released from the cells and purified using the REAL Prep 96  
5 plasmid kit (QIAGEN). This kit enabled the simultaneous purification of 96 samples in a  
96-well block using multi-channel reagent dispensers. The recommended protocol was  
employed except for the following changes: 1) the bacteria were cultured in 1 ml of sterile  
Terrific Broth (LifeTechnologies) with carbenicillin at 25 mg/L and glycerol at 0.4%; 2)  
after inoculation, the cultures were incubated for 19 hours and at the end of incubation, the  
10 cells were lysed with 0.3 ml of lysis buffer; and 3) following isopropanol precipitation, the  
plasmid DNA pellet was resuspended in 0.1 ml of distilled water. After the last step in the  
protocol, samples were transferred to a 96-well block for storage at 4° C.

The cDNAs were prepared and sequenced by the method of Sanger et al. (1975, J.  
Mol. Biol. 94:441f), using a Hamilton Micro Lab 2200 (Hamilton, Reno, NV) in  
15 combination with Peltier Thermal Cyclers (PTC200 from MJ Research, Watertown, MA)  
and Applied Biosystems 373 and 377 DNA Sequencing Systems.

## III. Selection, Assembly, and Characterization of Sequences

The sequences used for coexpression analysis were assembled from EST  
20 sequences, 5' and 3' longread sequences, and full length coding sequences. Selected  
assembled sequences were expressed in at least three cDNA libraries.

The assembly process is described as follows. EST sequence chromatograms were  
processed and verified. Quality scores were obtained using PHRED (Ewing, B. et al.  
(1998) Genome Res. 8:175-185; Ewing, B. and P. Green (1998) Genome Res. 8:186-194).  
25 Then the edited sequences were loaded into a relational database management system  
(RDBMS). The EST sequences were clustered into an initial set of bins using BLAST  
with a product score of 50. All clusters of two or more sequences were created as bins.  
The overlapping sequences represented in a bin correspond to the sequence of a  
transcribed gene.

30 Assembly of the component sequences within each bin was performed using a  
modification of Phrap, a publicly available program for assembling DNA fragments  
(Green, P., University of Washington, Seattle, WA). Bins that showed 82% identity from a



local pair-wise alignment between any of the consensus sequences were merged.

Bins were annotated by screening the consensus sequence in each bin against public databases, such as gbpri and genpept from NCBI. The annotation process involved a FASTn screen against the gbpri database in GenBank. Those hits with a percent identity  
 5 of greater than or equal to 70% and an alignment length of greater than or equal to 100 base pairs were recorded as homolog hits. The residual unannotated sequences were screened by FASTx against genpept. Those hits with an E value of less than or equal to  $10^{-8}$  are recorded as homolog hits.

Sequences were then reclustered using BLASTn and Cross-Match, a program for  
 10 rapid protein and nucleic acid sequence comparison and database search (Green, P., University of Washington, Seattle, WA), sequentially. Any BLAST alignment between a sequence and a consensus sequence with a score greater than 150 was realigned using cross-match. The sequence was added to the bin whose consensus sequence gave the highest Smith-Waterman score amongst local alignments with at least 82% identity. Non-  
 15 matching sequences created new bins. The assembly and consensus generation processes were performed for the new bins.

#### IV. Co-expression Analyses of Known Prostate Cancer-Specific Genes

Five known prostate cancer-specific genes were selected to test the validity of the  
 20 coexpression analysis method of the present invention in identifying genes that are closely associated with prostate cancer. The five known genes were prostate-specific antigen, glandular kallikrein, prostate seminal protein, prostatic acid phosphatase, and prostate transglutaminase. As shown, the method successfully identified the strong association of the known genes among themselves, indicating that the coexpression analysis method of  
 25 the present invention was effective in identifying genes that are closely associated with prostate cancer.

Table 4 shows the top ten genes that were most closely associated with a known prostate cancer-specific gene. These genes are presented along with their p-values. The column headings have the following meanings:

30

P-value	The probability that the observed number of
co-occurrences is due to chance using the Fisher	

exact method.

Co-expressed Gene A gene that shows significant co-expression with the target.

No. Occur. The number of libraries in which the associated gene occurs.

No. Co-occur The number of libraries in which both the target gene and the co-expressed gene occur.

No. Target Only The number of libraries in which **only** the target gene occurs.

No. Gene Only The number of libraries in which **only** the associated gene occurs.

No. Neither Occur The number of libraries in which neither the target gene nor the associated gene occur.

15

Table 4. Co-expression results for prostate-specific antigen

P-value	Co-expressed Gene	No. Occur	No. Co-occur	No. Target Only	No. Gene Only	No. Neither Occur
1.53E-31	Glandular kallikrein	29	26	12	3	481
1.65E-26	1816556	24	22	16	2	482
7.48E-25	1864683	40	26	12	14	470
8.12E-25	1344875	29	23	15	6	478
3.38E-24	Prostate seminal protein	30	23	15	7	477
1.89E-23	Prostatic acid phosphatase	35	24	14	11	473
6.87E-18	1651189	28	19	19	9	475
9.01E-18	Prostate trans-glutaminase	14	14	24	0	484
4.61E-14	1646118	93	27	11	66	418
1.58E-13	Human neuropeptide Y (NPY) mRNA	27	16	22	11	473

As a target, prostate-specific antigen occurred in 38 of 522 cDNA libraries studied, and showed strong coexpression with glandular kallikrein, prostate seminal protein, prostatic acid phosphatase, and prostate transglutaminase. The target also showed strong association with the human neuropeptide tyrosine (NPY) mRNA. In addition, four of the  
5 top ten genes that showed strong association with the target were novel Incyte assembled genes: 1816556, 1864683, 1344875, 1651189, and 1646118. These results are shown in Table 4 with association probability in the range of 1.58E-13 to 1.53E-31.

Similar results were observed when the other four known prostate cancer-specific genes, glandular kallikrein, prostate seminal protein, prostatic acid phosphatase, and  
10 prostate transglutaminase, were taken as target genes. These target genes were also found to be strongly associated with several known genes, many of which are cancer-related. The cancer-related genes are human neuropeptide tyrosine (NPY), human serine protease encoded by TMPRSS2, sorbitol dehydrogenase isozyme, human Zn-alpha-2-glycoprotein, and MAT-8.

NPY was first isolated from a human pheochromocytoma tumor (Minth, C. D. et al. (1984) Proc Natl Acad Sci 81(14): 4577-4581) and was reported to be involved in prostate cancer (Rogatnick, L. A. et al. (1990) Proc West Pharmacol Soc 33: 47-53; Mack, D., G. et al. (1997) Eur J Cancer 33:317-318). The TMPRSS2 gene was identified as a gene that encodes a serine protease domain specific for cleavage at Arg or Lys residues  
20 (Paoloni-Giacobino, A. et al. (1997) Genomics 44:309-320). The protease activity of TMPRSS2 is similar to that of PSA and kallikrein, both human prostate cancer-specific genes. Sorbitol dehydrogenase isozyme has been used as a marker for male reproductive tissue, including the prostate (Holmes, R. S. et al. (1978) J Exp Zool 206: 279-88). Significant activity of the enzyme accompanies damage to reproductive tissue.

Zn-alpha-2-glycoprotein is a secreted protein identified in hormone-responsive breast carcinomas (Freije, J. P. et al. (1993) Genomics 18:575-87) and was proposed as a marker for breast carcinomas (Lopez-Boado, Y. S. et al. (1994) Breast Cancer Res Treat 29: 247-58). It was shown to be prognostic for a 5-year breast cancer survival (Hurlimann, J. and G. van Melle (1991) Am J Clin Pathol 95:835-43) and was reported to show  
30 differential expression in prostate carcinoma (Gagnon, S. et al. (1990) Am J Pathol 136: 1147-52). MAT-8 was first identified in murine breast tumors and subsequently in primary human breast tumors and cell lines (Morrison, B. W. et al. (1995) J Biol Chem

270: 2176-82). It was shown to be a marker for progression of breast cancer (Schiemann, S. et al. (1998) Clin Exp Metastasis 16:129-39). A relation to prostate cancer has not been previously reported.

## 5 V. Identification of Novel Prostate Cancer-Associated Genes

Using the coexpression analysis, we have identified eight novel genes that show strong association with prostate cancer from a total of 41,419 assembled gene sequences. The degree of association was measured by probability values and has a cutoff of p value less than 0.00001. This was followed by annotation and literature searches to insure that  
 10 the genes that passed the probability test have strong association with known prostate cancer-specific genes. This process was reiterated so that the initial 41419 genes were reduced to the final eight prostate cancer-associated genes. Details of identification for the eight novel prostate cancer-associated genes are presented in Tables 5 to 12. These tables show the ten genes that were most closely associated for each target novel gene as  
 15 measured by coexpression using the Fisher exact test. The column headings have the same meanings as in Example IV.

Table 5. Co-expression results for Incyte gene 842349

	P-value	Co-expressed Gene	No. Occur	No. Co-occur	No. Target Only	No. Gene Only	No. Neither Occur
20	4.23E-11	Glandular kallikrein	29	17	38	12	455
	1.29E-09	Prostate seminal protein	30	16	39	14	453
	3.62E-09	1816556	24	14	41	10	457
	8.56E-09	1344875	29	15	40	14	453
	1.09E-08	TMPRSS2-encoded serine protease	74	24	31	50	417
25	1.19E-08	Prostate-specific antigen	38	17	38	21	446

2.45E-08	Prostatic acid phosphatase	35	16	39	19	448
2.91E-08	Human neuropeptide Y (NPY) mRNA	27	14	41	13	454
2.92E-08	1697453	96	27	28	69	398
3.14E-08	1864683	40	17	38	23	444

5

Incyte gene 842349 occurred in 55 of 522 cDNA libraries studied and showed strong co-expression with several of the known prostate cancer-specific genes, including glandular kallikrein, prostate seminal protein, prostate-specific antigen, and prostatic acid phosphatase, as shown in Table 9. 842349 also showed strong association with a human

10 TMPRSS2-encoded serine protease. The serine protease was shown to be strongly associated with prostate cancer-specific prostatic acid phosphatase in Example IV. Further, 842349 showed strong association with four novel Incyte genes, 1816556, 1344875, 1697453, and 1864683. These results are consistent with the notion that 842349 is associated with prostate cancer; and 842349 may be functionally or regulatorily

15 associated with at least four novel Incyte genes.

Table 6. Co-expression results for 1682557

	P-value	Co-expressed Gene	No. Occur	No. Co-occur	No. Target Only	No. Gene Only	No. Neither Occur
	1.34E-07	1816556	24	5	0	19	498
20	3.40E-07	Human nicotinic acetylcholine receptor A	10	4	1	6	511
	3.75E-07	Glandular kallikrein	29	5	0	24	493
	3.75E-07	1344875	29	5	0	24	493
	1.02E-06	Prostatic acid phosphatase	35	5	0	30	487

1.58E-06	Prostate-specific antigen	38	5	0	33	484
2.08E-06	1864683	40	5	0	35	482
4.22E-06	1685804	5	3	2	2	515
9.53E-06	3096181	21	4	1	17	500
2.34E-05	1794279	8	3	2	5	512

Incyte gene 1682557 occurred in 5 of 522 cDNA libraries studied and showed strong coexpression with several of the known prostate cancer-specific genes, such as glandular kallikrein, prostatic acid phosphatase, and prostate-specific antigen, as shown in Table 10. 1682557 also exhibited strong coexpression with the human nicotinic acetylcholine receptor A, a neurotransmitter which is a ligand-gated cation channel and causes rapid depolarization in postsynaptic cells. Further, Table 10 shows that 842349 has strong association with five novel Incyte genes, 1816556, 1344875, 1697453, 1864683, and 1794279. These results are consistent with the notion that 1682557 is associated with prostate cancer; and 1682557 may be functionally or regulatorily associated with at least five novel Incyte genes.

Table 7. Co-expression results for 1816556

P-value	Co-expressed Gene	No. Occur	No. Co-Occur	No. Target Only	No. Gene Only	No. Neither Occur
1.20E-30	Glandular kallikrein	29	22	2	7	491
1.11E-27	Prostatic acid phosphatase	35	22	2	13	485
1.65E-26	Prostate-specific antigen	38	22	2	16	482
1.55E-25	1344875	29	20	4	9	489
1.55E-23	1864683	40	21	3	19	479
8.19E-23	Prostate seminal protein	30	19	5	11	487

1.03E-16	Human gene for ZN-alpha-2-glycoprotein	86	22	2	64	434
3.13E-16	Prostate transglutaminase	14	12	12	2	496
1.08E-15	1651189	28	15	9	13	485
4.81E-15	2819055	44	17	7	27	471

5

Incye gene 1816556 occurred in 24 of 522 cDNA libraries studied and showed strong co-expression with several of the known prostate cancer-specific genes, such as glandular kallikrein, prostatic acid phosphatase, and prostate-specific antigen, prostate seminal protein, and prostate transglutaminase, as shown in Table 11. 1816556 also exhibited strong association with a human gene for ZN-alpha-2-glycoprotein which was shown in Example IV to be strongly associated with a prostate cancer-specific gene encoding prostatic transglutaminase. Further, 1816556 showed strong co-expression with four novel Incyte genes, 1344875, 1864683, 1651189, and 2819055. These results are consistent with the notion that 1816556 is associated with prostate cancer; and 1816556 may be functionally or regulatorily associated with at least four novel Incyte genes.

10  
15

Table 8. Co-expression results for 1864683

P-value	Co-expressed Gene	No. Occur	No. Co-Occur	No. Target Only	No. Gene Only	No. Neither Occur
7.48E-25	Prostate-specific antigen	38	26	14	12	470
4.55E-24	1344875	29	23	17	6	476
4.55E-24	Glandular kallikrein	29	23	17	6	476
1.55E-23	1816556	24	21	19	3	479
1.43E-21	Prostate seminal protein	30	22	18	8	474

20

6.52E-21	Prostatic acid phosphatase	35	23	17	12	470
5.43E-15	Prostate trans-glutaminase	14	13	27	1	481
6.38E-15	TMPRSS2-encoded serine protease	74	26	14	48	434
3.41E-14	1651189	28	17	23	11	471
1.47E-13	2819055	44	20	20	24	458

Incyte gene 1864683 occurred in 40 of 522 cDNA libraries studied and showed strong co-expression with several of the known prostate cancer-specific genes, such as prostate-specific antigen, glandular kallikrein, prostate seminal protein, prostatic acid phosphatase, and prostate transglutaminase, as shown in Table 8. 1864683 also exhibited strong association with a human TMPRSS2-encoded serine protease shown in Example IV to be strongly associated with prostate cancer-specific gene encoding prostatic acid phosphatase. Further, 1864683 showed strong coexpression with four novel Incyte genes, 1344875, 1816556, 1651189, and 2819055. These results are consistent with the notion that 1864683 is associated with prostate cancer; and 1864683 may be functionally or regulatorily associated with at least four novel Incyte genes.

Table 9. Co-expression results for 2187866

P-value	Co-expressed Gene	No. Occur	No. Co-Occur	No. Target Only	No. Gene Only	No. Neither Occur
9.11E-11	Prostatic acid phosphatase	35	9	1	26	486
2.42E-10	1816556	24	8	2	16	496
1.38E-09	Glandular kallikrein	29	8	2	21	491
1.38E-09	1344875	29	8	2	21	491



	1.53E-08	Prostate-specific antigen	38	8	2	30	482
	2.38E-08	1864683	40	8	2	32	480
	3.69E-08	Human LPAP gene	98	10	0	88	424
	5.41E-08	2819055	44	8	2	36	476
5	1.32E-07	TMPRSS2-encoded serine protease	74	9	1	65	447
	2.34E-06	843197	7	4	6	3	509

Incyte gene 2187866 occurred in 10 of 522 cDNA libraries and showed strong association with several of the known prostate cancer-specific genes, such as prostatic acid phosphatase, glandular kallikrein, and prostate-specific antigen, as shown in Table 13. 2187866 also exhibited strong association with a human lymphocyte phosphatase-associated phosphoprotein (LPAP) gene and a human TMPRSS2-encoded serine protease. LPAP is a 32 kDa protein that non-covalently binds tyrosine phosphatase CD45 (Bruyns, E., et al. (1998) Int Immunol 10: 185-94; Bruyns, E., A et al. (1996) Genomics 38: 79-83). As specified in Example IV, TMPRSS2-encoded serine protease was associated with a prostate cancer-specific gene encoding prostatic acid phosphatase. Further, 2187866 exhibited association with five novel Incyte genes, 1816556, 1344875, 1864683, 2819055, and 843197. These results are consistent with the notion that 2187866 is associated with prostate cancer; and 2187866 may be functionally or regulatorily associated with at least five novel Incyte genes.

Table 10. Co-expression results for 3096181

	P-value	Co-expressed Gene	No. Occur	No. Co-Occur	No. Target Only	No. Gene Only	No. Neither Occur
25	3.69E-13	1344875	29	13	8	16	485
	4.96E-10	Glandular kallikrein	29	11	10	18	483

	6.84E-10	Prostate-specific antigen	38	12	9	26	475
	8.49E-10	Human gene for ZN-alpha-2-glycoprotein	86	16	5	70	431
	1.38E-09	1816556	24	10	11	14	487
	5.35E-09	Prostatic acid phosphatase	35	11	10	24	477
5	1.87E-08	Prostate seminal protein	30	10	11	20	481
	2.70E-08	1864683	40	11	10	29	472
	1.26E-07	Human T-cell receptor gamma chain	27	9	12	18	483
	1.67E-07	Prostate transglutaminase	14	7	14	7	494

10

Incyte gene 3096181 occurred in 21 of 522 libraries studied and showed strong co-expression with several of the known prostate cancer-specific genes, such as glandular kallikrein, prostate-specific antigen, prostatic acid phosphatase, prostate seminal protein, and prostate transglutaminase, as shown in Table 14. 3096181 also exhibited strong

15 coexpression with a human gene for ZN-alpha-2-glycoprotein and human T-cell receptor gamma chain. As specified in Example IV, the human gene for ZN-alpha-2-glycoprotein was associated with a prostate cancer-specific gene encoding prostatic transglutaminase. ZN-alpha-2-glycoprotein itself was identified in hormone-responsive breast carcinomas (Freije et al., supra). Human T-cell receptor gamma/delta is expressed in tumor-infiltrating

20 lymphocytes in breast carcinoma patients (Alam, S.M. et al. (1992) Immunol Lett 31:279-283) and in malignant lymphoma presenting with hepatosplenic disease (Farcet, J. P. et al. (1990) Blood 75: 2213-2219). T-cell receptor gamma positive cells are over-expressed in cancer patients (Seki, S. et al. (1990) J Clin Invest 86: 409-15). Further, 3096181 exhibited coexpression with three novel Incyte genes, 1344875,

25 1816556, and 1864683. These results are consistent with the notion that 3096181 is

associated with prostate cancer; and 3096181 may be functionally or regulationally associated with at least three novel Incyte genes.

Table 11. Co-expression results for 3360806

5	P-value	Co-expressed Gene	No. Occur	No. Co-Occur	No. Target Only	No. Gene Only	No. Neither Occur
	1.98E-11	Prostate-specific antigen	38	16	18	22	466
	6.51E-10	1651189	28	13	21	15	473
	8.24E-10	1864683	40	15	19	25	463
	1.88E-08	1344875	29	12	22	17	471
10	1.88E-08	Glandular kallikrein	29	12	22	17	471
	3.00E-08	Prostate seminal protein	30	12	22	18	470
	5.10E-08	Human gene for ZN-alpha-2-glycoprotein	86	19	15	67	421
	3.19E-07	Prostate trans-glutaminase	14	8	26	6	482
	3.66E-07	1816556	24	10	24	14	474
15	6.56E-07	1685861	133	22	12	111	377

Incyte gene 3360806 occurred in 34 of 522 cDNA libraries and showed strong co-expression with several of the known prostate cancer-specific genes, such as prostate-specific antigen, glandular kallikrein, prostate seminal protein, and prostate  
 20 transglutaminase, as shown in Table 15. 3360806 also exhibited strong co-expression with a human gene for ZN-alpha-2-glycoprotein shown in Example IV to be associated with a prostate cancer-specific gene encoding prostatic transglutaminase.  
 ZN-alpha-2-glycoprotein itself was also found in hormone-responsive breast carcinomas (Freije et al., *supra*). Further, 3360806 showed association with five novel Incyte genes,  
 25 1651189, 1864683, 1344875, 1816556, and 1685861. These results are consistent with the

notion that 3360806 is associated with prostate cancer; and 3360806 may be functionally or regulationally associated with at least four novel Incyte genes.

Table 12. Co-expression results for 3458076

5	P-value	Co-expressed Gene	No. Occur	No. Co-Occur	No. Target Only	No. Gene Only	No. Neither Occur
	3.35E-08	1816556	24	6	1	18	497
	1.17E-07	1344875	29	6	1	23	492
	1.17E-07	Glandular kallikrein	29	6	1	23	492
	1.46E-07	Prostate seminal protein	30	6	1	24	491
10	3.96E-07	Prostatic acid phosphatase	35	6	1	29	486
	6.70E-07	Prostate-specific antigen	38	6	1	32	483
	9.29E-07	1864683	40	6	1	34	481
	6.04E-06	1651189	28	5	2	23	492
	1.09E-05	Prostate trans-glutaminase	14	4	3	10	505
15	1.22E-05	Human polymorphic CA dinucleotide repeat	32	5	2	27	488

Incyte gene 3458076 occurred in 7 of 522 cDNA libraries and showed strong co-expression with several of the known prostate cancer-specific genes, such as glandular kallikrein, prostate seminal protein, prostatic acid phosphatase, prostate-specific antigen, and prostate transglutaminase, as shown in Table 11. 3458076 also exhibited association with a human dinucleotide repeat flanking region, a region that flanks a polymorphic CA micro satellite repeat from the long arm of chromosome 1 (Raymond, M. H., et al. (1987) GI 2182124, GenBank). Genes in this region have not been characterized. Further,

3458076 showed coexpression with four novel Incyte genes, 1816556, 1344875, 1864683, and 1651189. These results are consistent with the notion that 3458076 is associated with prostate cancer; and 3458076 may be functionally or regulationally associated with at least four novel Incyte genes.

5

## VI. Novel Prostate Cancer-Associated Genes

Eight novel Incyte genes were identified from the data shown in Tables 5 to 12 to be associated with prostate cancer.

Nucleic acids comprising the consensus sequences of SEQ ID NOs: 1-10 of the present invention were first identified from Incyte Clones 842349, 1682557, 1816556, 1864683, 2187866, 3096181, 3360806, and 3458076, respectively, and assembled according to Example III. BLAST and other motif searches were performed for SEQ ID NOs: 1-8 according to Example VII. The sequences were translated and sequence identity was sought with known sequence. Of interest, the amino acid sequence encoded by SEQ ID NO: 1 from about nucleotide 195 to about nucleotide 446 showed 58% sequence identity with a subunit of a mouse RNA polymerase I, PRA16 (GI 1778684); and the amino acid sequence encoded SEQ ID NO:3 from about 185 to about 825 showed about 76% sequence identity with a Sus scrofa enamel matrix serine protease (GI 2737921). The protease activity of the enamel matrix serine protease is consistent with that of PSA and kallikrein, two of the known human prostate cancer-specific genes. SEQ ID NO: 9 is an amino acid sequence coded by SEQ ID NO: 4. SEQ ID NO: 9 is 231 amino acids in length. Residue 188 to residue 209 encompass a potential transmembrane domain, and residue 1 to residue 47 is a potential signal peptide sequence. SEQ ID NO: 9 also has two potential casein kinase II phosphorylation sites at S100 and S142; one potential protein kinase C phosphorylation site at S147; and a potential cell attachment sequence encompassing residues R93GD which interacts with a cell surface receptor. SEQ ID NO: 10 is an amino acid sequence coded by SEQ ID NO: 8. SEQ ID NO: 10 is 162 amino acids in length. The fragment from residue 83 to residue 99 resembles a potential BLOCK signature of Ly-6/u-PAR, a family of GPI-linked cell-surface glycoproteins. SEQ ID NO: 10 also has one potential N-glycosylation site at N4; one potential cAMP- and cGMP-dependent protein kinase phosphorylation site at T48; and three potential protein kinase C phosphorylation sites at T25, T34, and S44.

## VII. Homology Searching for Prostate Cancer-Associated Genes and the Proteins Encoded by the Genes

Polynucleotide sequences, SEQ ID NOs: 1-8, and polypeptide sequences, SEQ ID  
5 NOs: 9-10, were queried against databases derived from sources such as GenBank and  
SwissProt. These databases, which contain previously identified and annotated sequences,  
were searched for regions of similarity using Basic Local Alignment Search Tool  
(BLAST; Altschul, S.F. et al. (1990) J. Mol. Biol. 215: 403-410) and Smith-Waterman  
alignment (Smith, T. et al. (1992) Protein Engineering 5:35-51). BLAST searched for  
10 matches and reported only those that satisfied the probability thresholds of  $10^{-25}$  or less for  
nucleotide sequences and  $10^{-8}$  or less for polypeptide sequences.

The polypeptide sequences were also analyzed for known motif patterns using  
MOTIFS, SPSCAN, BLIMPS, and Hidden Markov Model (HMM)-based protocols.  
MOTIFS (Genetics Computer Group, Madison, WI) searches polypeptide sequences for  
15 patterns that match those defined in the Prosite Dictionary of Protein Sites and Patterns  
(Bairoch, A. et al. (1997) Nucleic Acids Res. 25:217-221), and displays the patterns found  
and their corresponding literature abstracts. SPSCAN (Genetics Computer Group,  
Madison, WI) searches for potential signal peptide sequences using a weighted matrix  
method (Nielsen, H. et al. (1997) Prot. Eng. 10: 1-6). Hits with a score of 5 or greater  
20 were considered. BLIMPS uses a weighted matrix analysis algorithm to search for  
sequence similarity between the polypeptide sequences and those contained in BLOCKS, a  
database consisting of short amino acid segments, or blocks, of 3-60 amino acids in length,  
compiled from the PROSITE database (Henikoff, S. and G. J. Henikoff (1991) Nucleic  
Acids Res. 19:6565-6572; Bairoch et al., supra), and those in PRINTS, a protein  
25 fingerprint database based on non-redundant sequences obtained from sources such as  
SwissProt, GenBank, PIR, and NRL-3D (Attwood, T. K. et al. (1997) J. Chem. Inf.  
Comput. Sci. 37:417-424). For the purposes of the present invention, the BLIMPS  
searches reported matches with a cutoff score of 1000 or greater and a cutoff probability  
value of  $1.0 \times 10^{-3}$ . HMM-based protocols were based on a probabilistic approach and  
30 searched for consensus primary structures of gene families in the protein sequences (Eddy,  
S.R. (1996) Cur. Opin. Str. Biol. 6:361-365; Sonnhammer, E.L.L. et al. (1997) Proteins  
28:405-420). More than 500 known protein families with cutoff scores ranging from 10 to

50 bits were selected for use in this invention.

### VIII. Extension of Polynucleotides

The initial primers were designed from the cDNA using OLIGO 4.06 (National  
5 Biosciences, Plymouth, MN), or another appropriate program, to be about 22 to 30  
nucleotides in length, to have a GC content of about 50% or more, and to anneal to the  
target sequence at temperatures of about 68°C to about 72°C. Any stretch of nucleotides  
which would result in hairpin structures and primer-primer dimerizations were avoided.  
Selected human cDNA libraries (GIBCO/BRL) were used to extend the sequence. If more  
10 than one extension is necessary or desired, additional sets of primers are designed to  
further extend the known region.

High fidelity amplification was obtained by following the instructions for the XL-  
PCR kit (Perkin Elmer) and thoroughly mixing the enzyme and reaction mix. PCR was  
performed using the Peltier Thermal Cycler (PTC200; M.J. Research, Watertown, MA),  
15 beginning with 40 pmol of each primer and the recommended concentrations of all other  
components of the kit.

A 5  $\mu$ l to 10  $\mu$ l aliquot of the reaction mixture was analyzed by electrophoresis on  
a low concentration (about 0.6% to 0.8%) agarose mini-gel to determine which reactions  
were successful in extending the sequence. Bands thought to contain the largest products  
20 were excised from the gel, purified using QIAQuick™ (QIAGEN), and trimmed of  
overhangs using Klenow enzyme to facilitate religation and cloning.

After ethanol precipitation, the products were redissolved in 13  $\mu$ l of ligation  
buffer, 1  $\mu$ l T4-DNA ligase (15 units) and 1  $\mu$ l T4 polynucleotide kinase were added, and  
the mixture was incubated at room temperature for 2 to 3 hours, or overnight at 16° C.  
25 Competent *E. coli* cells (in 40  $\mu$ l of appropriate media) were transformed with 3  $\mu$ l of  
ligation mixture and cultured in 80  $\mu$ l of SOC medium. After incubation for one hour at  
37° C, the *E. coli* mixture was plated on Luria Bertani (LB) agar containing 2x Carb. The  
following day, several colonies were randomly picked from each plate and cultured in 150  
 $\mu$ l of liquid LB/2x Carb medium placed in an individual well of an appropriate  
30 commercially-available sterile 96-well microtiter plate.

**IX. Labeling and Use of Individual Hybridization Probes**

Oligonucleotides are designed using state-of-the-art software such as OLIGO 4.06 (National Biosciences) and labeled by combining 50 pmol of each oligomer, 250  $\mu$ Ci of [ $\gamma$ - $^{32}$ P] adenosine triphosphate (Amersham, Chicago, IL), and T4 polynucleotide kinase (DuPont NEN®, Boston, MA). The labeled oligonucleotides are substantially purified using a Sephadex G-25 superfine resin column (Pharmacia & Upjohn, Kalamazoo, MI). An aliquot containing  $10^7$  counts per minute of the labeled probe is used in a typical membrane-based hybridization analysis of human genomic DNA digested with one of the following endonucleases: Ase I, Bgl II, Eco RI, Pst I, Xba I, or Pvu II (DuPont NEN, Boston, MA).

The DNA from each digest is fractionated on a 0.7 percent agarose gel and transferred to nylon membranes (Nytran Plus, Schleicher & Schuell, Durham, NH). Hybridization is carried out for 16 hours at 40°C. To remove nonspecific signals, blots are sequentially washed at room temperature under increasingly stringent conditions up to 0.1 x saline sodium citrate and 0.5% sodium dodecyl sulfate. After XOMAT AR™ film (Kodak, Rochester, NY) is exposed to the blots to film for several hours, hybridization patterns are compared visually.



What is claimed is:

1. A method for identifying biomolecules useful in the diagnosis or treatment of a disease associated with cell proliferation, said method comprising:
  - 5 a) examining expression patterns of polynucleotides that are expressed in a plurality of cDNA libraries, said expressed polynucleotides comprising one or more prostate cancer-specific genes and one or more genes of unknown function; and
  - b) comparing the expression patterns of said prostate cancer-specific genes with the expression patterns of the genes of unknown function to identify a subset of the genes of  
10 unknown function which have similar expression patterns to those of prostate cancer-specific genes.
- 2 The method of claim 1, wherein said polynucleotides are selected from the group consisting of expressed sequence tags (ESTs), assembled sequences, full length  
15 gene coding sequences, 5' untranslated regions and 3' untranslated regions.
3. The method of claim 1, wherein said prostate cancer-specific genes are selected from the group consisting of prostate-specific antigen, prostatic acid phosphatase, kallikrein, seminal plasma protein and prostate-specific transglutaminase.  
20
4. The method of claim 1, wherein said comparing comprises
  - a) generating an occurrence data vector for each expressed polynucleotide;
  - b) analyzing vectors for two or more expressed polynucleotides to determine a coexpression probability; and
  - 25 c) determining whether the coexpression probability for said two or more expressed polynucleotides is less than a specified coexpression probability threshold.
5. The method of claim 1, further comprising the step of translating said subset of genes of unknown function to generate corresponding polypeptides.  
30
6. A polynucleotide identified by the method of claim 1.

7. A polypeptide identified by the method of claim 5.

8. A substantially purified biomolecule for use in the diagnosis or treatment of a disease associated with cell proliferation, said biomolecule selected from the group
- 5 consisting of:
- (A) a polynucleotide selected from the group consisting of SEQ ID NOs: 1-8;
  - (B) a polynucleotide which encodes a polypeptide selected from the group consisting of SEQ ID NOs: 9-10;
  - (C) a polynucleotide having at least 70% identity to the polynucleotide of (A) or
  - 10 (B);
  - (D) a polynucleotide which is complementary to the polynucleotide of (A), (B), or (C);
  - (E) a polynucleotide comprising at least 18 sequential nucleotides of the polynucleotide of (A), (B), (C), or (D);
  - 15 (F) a polypeptide selected from the group consisting of SEQ ID NOs: 9-10;
  - (G) a polypeptide having at least 85% identity to the polypeptide of (F); and
  - (H) a polypeptide comprising at least 6 sequential amino acids of the polypeptide of (F) or (G).

- 20 9. The substantially purified biomolecule of claim 8, comprising a polynucleotide sequence selected from the group consisting of:
- (A) a polynucleotide selected from the group consisting of SEQ ID NOs: 1-8;
  - (B) a polynucleotide which encodes a polypeptide selected from the group consisting of SEQ ID NOs: 9-10;
  - 25 (C) a polynucleotide having at least 70% identity to the polynucleotide of (A) or (B);
  - (D) a polynucleotide which is complementary to the polynucleotide of (A), (B), or (C);
  - (E) a polynucleotide comprising at least 18 sequential nucleotides of the
  - 30 polynucleotide of (A), (B), (C), or (D); and
  - (F) a polynucleotide which hybridizes under stringent conditions to the polynucleotide of (A), (B), (C), (D), or (E).

10. The substantially purified biomolecule of claim 8, comprising a polypeptide sequence selected from the group consisting of:

- (A) a polypeptide selected from the group consisting of SEQ ID NOs:9-10;
- (B) a polypeptide having at least 85% identity to the polypeptide of (A); and
- 5 (C) a polypeptide comprising at least 6 sequential amino acids of the polypeptide of (A) or (B).

11. An expression vector comprising the polynucleotide of claim 9.

10 12. A host cell comprising the expression vector of claim 11.

13. A method for producing a polypeptide of claim 10, the method comprising the steps of:

- a) culturing the host cell of claim 12 under conditions suitable for the
- 15 expression of the polypeptide; and
- b) recovering the polypeptide from the host cell culture.

14. A pharmaceutical composition comprising the biomolecule of claim 8 in conjunction with a suitable pharmaceutical carrier.

20

15. An antibody which specifically binds to the polypeptide of claim 10.

16. A method for detecting a target polynucleotide in a biological sample, the method comprising the steps of:

- 25 (a) hybridizing the polynucleotide of claim 9 to the target polynucleotide to form a hybridization complex; and
- (b) detecting the hybridization complex, wherein the presence of the hybridization complex correlates with the presence of the target polynucleotide.

30 17. A method for identifying biomolecules useful in the diagnosis or treatment of a disease, said method comprising:

- a) examining expression patterns of a plurality of biomolecules that are expressed

in a plurality of cDNA libraries, said expressed biomolecules comprising one or more disease-specific biomolecules and one or more biomolecules of unknown function; and

- b) comparing the expression patterns of said disease-specific biomolecules with the expression patterns of the biomolecules of unknown function to identify a subset of the
- 5 biomolecules of unknown function which have similar expression patterns to those of the disease-specific biomolecules.

18. The method of claim 17, wherein said comparing comprises

- a) generating an occurrence data vector for each expressed biomolecule;
- 10 b) analyzing vectors for two or more expressed biomolecules to determine a coexpression probability; and
- c) determining whether the coexpression probability for said two or more expressed biomolecules is less than a specified coexpression probability threshold.

15 19. A polynucleotide identified by the method of claim 17.

20. A polypeptide identified by the method of claim 17.

## SEQUENCE LISTING

<110> INCYTE PHARMACEUTICALS, INC.

WALKER, Michael G.

VOLKMUTH, Wayne

KLINGLER, Tod M.

SPRINZAK, Einat A.

<120> PROSTATE CANCER-ASSOCIATED GENES

<130> PB-0002 PCT

<140> To Be Assigned

<141> Herewith

<150> 09/102,615

<151> 1998-06-22

<160> 10

<170> PERL Program

<210> 1

<211> 2215

<212> DNA

<213> Homo sapiens

<220>

<221> misc\_feature

<223> Incyte clone 842349

<400> 1

angtcatnta ttatgaagat nacacaggtt cttgcagaga ttccaacagc tgcaaggggt 60  
caacagctag aatcangcct ttctgaagga cagagtatgc tgtaaccaca aacttctaata 120  
tctgggttct gcnccatcan gaanagaata tcctacagga cagttctcct tgtatactgc 180  
ataaaggact anaatgtgga ttcatttctg cttgcttnt gatccttata gtcctttatg 240  
ctggcctcaa acttgtaag cacatgttg cagacattna tgagctcatt caggcctctc 300  
tganatggct caacagctgg aatggtaacct cngtctgaa tgcgtaaatn aagtttgctc 360  
tcngaaggat gggntcggag tggaaccaca aaattncnct tncggngtc gtcatgatct 420  
attatcatgt actttctaac tgagccctct attttcttta ttttaataat atttctcccc 480  
acttgagaat cacttgtag ttcttgtag gaattcagtt gggcaatgat aacttttatg 540  
ggcaaaaaca ttctattata gtgaacaaat gaaaataaca gcgtattttc aatattttct 600  
tattccttaa attccactct ttaacacta tgcttaacca cttaatgtga tgaaatattc 660  
ctaaaagtta aatgactatt aaagcatata ttgttgcatg tatatatata gtagccgata 720  
ctctaaataa aaataccact gttacagata aatggggcct ttaaaaaat gaaaaacaaa 780  
cttggtgaaaa tgtataaaa atgcatctgt tgtttcaaat ggcactatct tcttttcagt 840  
actacaaaaa cagaataatt ttgaagttt agaataaatg taatatattt actataattc 900  
taaagtgtta aatgcttttc taaaaatgca aaactatgat gtttagttgc tttattttac 960  
ctctatgtga ttatttttct taattgttat tttttataat cattattttt ctgaaccatt 1020  
cttctggcct cagaagtagg actgaattct actattgcta ggtgtgagaa agtgggtggtg 1080  
agaaccttag agcagtgagg atttgctacc tgggtctgtgt tttgagaagt gccccttaga 1140  
aagttaaaag aatgtagaaa agatactcag tcttaatcct atgcaaaaaa aaaaaatcaa 1200  
gtaattgttt tcctatgagg aaaataacca tgagctgtat catgctactt agcttttatg 1260  
taaataattt ttatgtctcc tctattaaga gtatttaaaa tcatatttaa atatgaattc 1320  
attcatgcta acattatttt tcaaaacata catggaaatt tagcccagat tgtctacata 1380  
taagggtttt atttgaattg taaaatattt aaaagtatga ataaaatata tttataggta 1440

```

tttatcagag atgattattt tgtgctacat acagggtggc taatgagctc tagtggttaa 1500
ctacctgatt aatttcttat aaagcagcat aaccttggct tgattaagga attctacttt 1560
caaaaattaa tctgataata gtaacaagggt atattatact ttcattacaa tcaaattata 1620
gaaattactt gtgtaaaagg gcttcaagaa tatatccaat ttttaaatat tttaatatat 1680
ctcctatctg ataacttaat tcttctaaat taccacttgc cattaagcta tttcataata 1740
aattctgtac agtttcccc aaaaaaaga gatttattta tgaaatattt aaagtttcta 1800
atgtggtatt taaataaagt aatcatnaat gnnataagtn aatatttatt taggaatact 1860
gtgaaacnct ganactaatt attnctcgtg tcagtcctat nnnantccct ggttttggga 1920
natnngnnaa nccagcctan aaatagnngt tgnmnaaatn anmttggtg aaannncatg 1980
nancctttaa annccntggg ttacnnttcn cntnggtatn tnaccanccc ccttccccac 2040
nnggtttaan aantttaatn aagtgttga aaatgggggc ctantcntna gcctttacaa 2100
cnttgtgtcc caatcanctg nggggtttna ancacaantt cccggatngg gtnnanttnc 2160
nnngggggnnc caactcgtgt nntgnccnc ncgacnnttg nacctcggnn aacna 2215

```

&lt;210&gt; 2

&lt;211&gt; 742

&lt;212&gt; DNA

&lt;213&gt; Homo sapiens

&lt;220&gt;

&lt;221&gt; misc\_feature

&lt;223&gt; Incyte clone 1682557

&lt;400&gt; 2

```

agcgtggggc tgetggactt ctgtgggcca ggtgtgcttc actccactgg aggccttgc 60
ctctgacctc ttccgggacc cggaccactg tcgccaggcc tactctgtct atgccttcat 120
gattagtctt gggggctgcc tgggctacct cctgcctgcc attgactggg acaccagtgc 180
cctggccccc tacctgggca cccaggagga gtgcctcttt ggcctgctca ccctcatctt 240
cctcacctgc gtagagccca cactgctggt ggctgaggag gcagcgctgg gcccaccga 300
gccagcagaa gggctgtcgg cccctctctt gtgcgccac tgctgtccat gccggggccc 360
cttggtcttc cggaacctgg gcgcctgct tccccggctg caccagtgt gctgcgcgat 420
gccccgcacc ctgcgcgggc tcttcgtggc tgagctgtgc agctggatgg cactcatgac 480
ctnatngnat tcaattnaac tttcaaagggt ttnttttaaa aangggattt ttgntggggg 540
gnaanngggc tnttaacann ggncnttgc ccanaannc ntnaannccg gggaaannnn 600
ngggncnng aaananaatt ttngnatnan aaggcntttt gnaatnggnc aanccttnggg 660
gttttttccc ngnaatnann gnantttccc cngntttttn ggtntntnnn ccaanngcc 720
ccgcntgggg gnaattantt tt 742

```

&lt;210&gt; 3

&lt;211&gt; 1040

&lt;212&gt; DNA

&lt;213&gt; Homo sapiens

&lt;220&gt;

&lt;221&gt; misc\_feature

&lt;223&gt; Incyte clone 1816556

&lt;400&gt; 3

```

ggggccctgg acggttttct ctatccactc agtgaatttg cagagggttg tntagacacc 60
tggcgcgcca acttgccac acggggcttt tccgaaagac acaaggccct gcaagtaccc 120
gttgacagatc aggggcccc cagagtcacc gttgcaggag tccttctggt cttgcctcc 180
gccggcgag gactgcagcc cgcgctcgca gccctggcag gcggcactgg tcatggaaaa 240
cgaattgttc tgctcgccgc acactgtttc cagaagttag tgacagctc ctacaccatc 300
gggctgggccc tgcacagtct tgaggccgac caagagccag ggagccagat ggtggaggcc 360
agcctctccg tacggcacc agagtacaac agacccttgc tcgctaacga cctcatgctc 420

```

```

atcaagtgg acgaatccgt gtccgagctt gacaccatcc ggagcatcag cattgcttcg 480
cagtgcctta ccgcggggaa ctcttgcttc gtttctggct ggggtctgct ggcgaacggc 540
agaatgccta ccgtgctgca gtgcgtgaac gtgtcggtgg tgtctgagga ggtctgcagt 600
aagctctatg acccgctgta ccacccagc atgttctgct ccggcggagg gcaagaccag 660
aaggactcct gcaacggtga ctctgggggg ccctgatctt gcaacgggta ctgacggggc 720
cttgtgtctt tcggaaaagc cccgtgtggc caagttggcg tgccagggtg ctacaccaac 780
ctctgcaaat tcaactgagt gatagagaaa accgtccagg ccagttaact ctggggactg 840
ggaacccatg aaattgacct ccaaatacat cctgcggaag gaattcagga atatctgttc 900
ccagccccct gtcctnagg ccaggagtc cacnncnaa aanantcnnt ncnctaaanc 960
nagggggtac anaatcccc aanaaaggan nntnncannn angacccang ggannnnnaa 1020
aaatttccnn naaaataggc                                     1040

```

&lt;210&gt; 4

&lt;211&gt; 2462

&lt;212&gt; DNA

&lt;213&gt; Homo sapiens

&lt;220&gt;

&lt;221&gt; misc\_feature

&lt;223&gt; Incyte clone 1864683

&lt;400&gt; 4

```

caacgacttt ccaaataatc tcaccagcgc cttccagctc aggcgtccta gaagcgtctt 60
gaagectatg gccagctgtc tttgtgttcc ctctcaccgc cctgtcctca cagctgagac 120
tcccaggaaa ccttcagact accttctctt gccttcagca aggggcgttg cccacattct 180
ctgaggggtca gtggaagaac ctgactccc attgctagag gtagaaaagg gaagggtgct 240
ggggagcagg gctggtccac agcaggtctc gtgcagcagg tactgtgtgt tccgccttct 300
catctccctg agactgctcc gaccttccc tcccaggctc tgtctgatgg cccctctccc 360
tctgcaggcg ttcggtggg cagcctgggg ctgttctctg agtgcgccat ctccctgggt 420
ttctctctgg tcatggaccg gctggtgcag cgattcggca ctcgagcagt ctatttggcc 480
agtgtggcag ctttccctgt ggtgcccgtt gccacatgcc tgtcccacag tgtggccgtg 540
gtgacagctt cagccgccct caccgggttc accttctcag cctgcagat cctgccctac 600
acactggcct ccctctacca ccgggagaag caggtgttcc tgcccaaata ccgaggggac 660
actggagggtg ctgacagtga ggacagcctg atgaccagct tcctgccagg ccctaagcct 720
ggagctccct tcctaatagg acacgtgggt gctggaggca gtggcctgct cccacctcca 780
ccgcgctctt ggggggcctc tgctcgtgat gtgtgggtgt gggtagagcc 840
accgaggcca ggggtgttcc gggccggggc atctgcctgg acctcgccat cctggatagt 900
gccttctctg tgcccagggt ggccccatcc ctgtttatgg gtccattgt ccagctcagc 960
cagtctgtca ctgcctatat ggtgtctgcc gcagcgtgg gtctggtcgc catttacttt 1020
gctacacagg tagtatttga caagagcgac ttggccaaat actcagcgtg gaaaacttcc 1080
agcacattgg ggtggagggg ctgcctcact ggggtcccag tccctgctcc tgttagcccc 1140
atggggctgc cgggctggcc gccagtttct gttgctgcca aagtaatgtg gctctctgct 1200
gccaccctgt gctgctgagg tgcgtactgc acagctgggg gctggggcgt cctctctctc 1260
tctccccagt ctctaggggt gcctgactgg aggccttcca aggggggttc agtctggact 1320
tatacaggga ggccagaagg gctccatgca ctggaatgcg gggactctgc aggtggatta 1380
cccaggctca gggttaacag ctagcctcct agttgagaca cacctagaga aggggttttg 1440
ggagctgaat aaactcagtc acctggtttc ccatctctaa gccccttaac ctgcagcttc 1500
gtttaatgta gctcttgcct gggagtttct aggatgaaac actcctccat gggatttgaa 1560
catatgaaag ttattttagt gggaagagtc ctgaggggca acacacaaga accaggtccc 1620
ctcagcccac agcactgtct ttttctgtat ccacccccct cttacctttt atcaggatgt 1680
ggcctgttgg tccttctgtt gccatcacag agacacaggc atttaaatat ttaacttatt 1740
tatttaacaa agtagaaggg aatccattgc tagcttttct gtgttggtgt ctaatatattg 1800
ggtagggtgg gggatcccca acaatcaggt cccctgagat agctggtcat tgggctgatc 1860
attgccagaa tcttcttctc ctggggtctg gcccccaaa atgcctaacc caggaccttg 1920
gaaattctac tcatcccaaa tgataattcc aaatgctgtt acccaagggt aggggtgttg 1980
aggaaggtag aggggtggggc ttcaggtctc aacggcttcc ctaaccaccc ctcttctctt 2040

```

```

ggcccagcct ggttccccc acttccactc cctctactc tctctaggac tgggctgatg 2100
aaggcactgc ccaaaatttc ccctaccccc aactttcccc tacccecaac tttccccacc 2160
agctccacaa cctgttttgg agctactgca ggaccagaag cacaaagtgc ggtttcccaa 2220
gcctttgtcc atctcagccc ccagagtata tctgtgcttg gggaaatctca cacagaaact 2280
caggagcacc cctgcctga gctaagggag gtcttatctc tcaggggggg ttttaagtgc 2340
gtttgcaata atgtcgtctt atttatttag cggggtgaat attttatact gtaagtgagc 2400
aatcagagta taatgtttat ggtgacaaaa ttaaaggctt tcttatatgt taaaaaaaaa 2460
aa 2462

```

&lt;210&gt; 5

&lt;211&gt; 1567

&lt;212&gt; DNA

&lt;213&gt; Homo sapiens

&lt;220&gt;

&lt;221&gt; misc\_feature

&lt;223&gt; Incyte clone 2187866

&lt;400&gt; 5

```

gcctcatccc tgggttgttg tgtggacatt gtgggtgtct ccacaggagc cccagggcca 60
cgaaagctgg ggtggcctct gcccttcttg gggttccttt tcctgcacag ctgctttctg 120
actccacca cagctgggag caggtgccgg agccccggcc tgcattggccc tgtgaaggcc 180
actctgggcg tttgggtggg cgtgagtgcc ttctctgct cccagcatgt ggttttctcc 240
gttggccgcc tcctggacct cctgggtgcct gacatcccag agtctgtgga gatcaaagtg 300
aagcgggagt actacctggc taagcaggca ctggctgaga atgaggttct ttttggaaacg 360
aacggaacaa aggatgagca gcccaggggc tcagagctca gctccactg gacacccttc 420
acggttccca aggccagcca gctgcagcag tgacgcctgg aaggacatct ggtggctcctt 480
aggggagtg gcccctctga gccctgcgag cagcgtcctt ttctcttcc ctcaggcagc 540
ggctgtgtga accgctggct gctgttgtgc ctcatctctg ggcacattgc ctgcttcccc 600
ccagcgccgg cttctctcct cagagcgcc gtcactccat ccccgccagg gagggaccgt 660
cagctcacaa ggccctcttt gtttctctg cccagacata agcccaagg gcccctgcac 720
ccaagggacc ctgtccctcg gtggcctccc caggcccttg gacacgacag ttctcctcag 780
gcaggtgggc tttgtgggtc tcgccgcccc tggccacatc gccctctcct cttacacctg 840
gtgaccttcg aatgtttcag agcgcagggc cgttctcct cgtgtcctct ggacccaccc 900
gccccttct gccctgtttg cgcagggaca tcacccacat gcccagctc tcggaccctg 960
cagctctgtg tcccaggcca cagcaaagg ctgttgaaac cctccctcca ttcccagtta 1020
tctgggtcct ctggattcct ctgtttcttg aatcaggctc tgctttcccc ctagccacta 1080
caggcagcct ctgacagtgc cgctttactt gcattctgca gcaattacat gtgtcctttt 1140
gatccttgcc caacttccct cctctccca gctcctggcc cctggcccag ggccctctt 1200
gctgttttta cctctgttcc ttggggccta gtaccagca agcaccaaa tgggggaggt 1260
tttgggatga gaggaggaaa cgtgtatacc tgtaacatct ggtggctctt ccccagaag 1320
tttgtgttca tacataattg tttccacgc tggatcataa tgtgacgtgc agttctgccc 1380
tgtgtgtggg agccacatga agcttccct ggctaacttg ctaccccgca gcaatcccag 1440
tgtggctgtc tgcttgctaa aaaatggatc tgtgtctatc tgtattgatg tccttgaggt 1500
tctacaagtg gaacttaagt gtcaaaaaga atatgtggtt tttaggggga tcctctagag 1560
tcgacct 1567

```

&lt;210&gt; 6

&lt;211&gt; 1354

&lt;212&gt; DNA

&lt;213&gt; Homo sapiens

&lt;220&gt;

&lt;221&gt; misc\_feature

&lt;223&gt; Incyte clone 3096181



&lt;400&gt; 6

```

taagaatgag tttcttactg aacaactttc taaaacgcaa attaaactca ataccttaaa 60
agataagttc cgtaagacaa gagatactct cagaaaaaag tcatcggcct tagaaactct 120
ccaaacgacc taagccaaac acagcagcaa ataaaggaaa tgaaagagat gtatgaaaat 180
gcagaagata aagtgaataa ttccactgga aagtggagct gtgtagaaga gaggatatgt 240
catctccaac atgaaaatcc gtgccttgaa cagcaactag atgatgttca tcagaaagag 300
gatcataaag agatagtaac taatatccaa agaggcctta ttgagagtag aaagaaagac 360
ctcatgctag aagagaaaag taagaagcta atgaatgaat gtgatcattt aaaagaaagt 420
ctctttcaat atgagagaga gaaagcagaa agagtagtgg ttgtgagaca acttcaacaa 480
gaagcggctg acagccctaa aaaaattaac tatgttagag tctccactgg aagggtatata 540
acgttgtcat attaatgttg ttgagacaca ggtcccaaag aagaaattat ttcaagtgga 600
aagtcaattt gatgatctta tggcggagaa ggaagctgta tcttcaaaat gtgtcaattt 660
ggctaaagag aatcaagttt ttcaacagga gttattatct atgaaaaaag cacaacagga 720
atgtgaaaaa tttgaggagg ataaaagatg ttggaagaag aaatattaaa tcttaagaca 780
catatggaaa acagtatggt agaacttagt aaactacaag aatataaatc agagctagat 840
gaaaaggcaa tgcaggcagt agaaaaatta gaagaaatcc atttacagga acaagcacia 900
tataaaaaac aattagagca gttaacaag gatataatac agcttcaact aataagaagg 960
aactcacact taaagatgtg gaatgtaaat tctacaaaat gaaaactgct tatgaagagg 1020
ttacaactga gttagaagaa tataaggaag cctttgcagc agcattgana gctaacaggt 1080
ccatgtctaa nanattaact aaatcgata aganaatagc aatgatcagt atcagctctt 1140
tatggngana gaggcgtga natattttct cagcactctt cctacnnggc gaggtcgaga 1200
gtcaccttgt gttgantatn cctacttngt ataggngctc agtcaganga tataatnctt 1260
ccaaatngcc cgtntggggg tectancttt caggngcctn cagagnttcc ataattnaac 1320
ngnccnaggn gtanctttga nctgagggcc nntt 1354

```

&lt;210&gt; 7

&lt;211&gt; 2426

&lt;212&gt; DNA

&lt;213&gt; Homo sapiens

&lt;220&gt;

&lt;221&gt; misc\_feature

&lt;223&gt; Incyte clone 3360806

&lt;400&gt; 7

```

ctcccttctt tctcaggggg tcccagcccc cgactagctt tgnccctaact ccttcatcaa 60
aaganncccc gccagcttcc cacacctcat acgcagccac atctgcccta ttctccatgc 120
tttccagctt gcctgccctt cctcatctct cctgcctgt gcagacctcc acccttcttt 180
cctccacccc tccatcccc aatgcttgta gaccttccat tcattccgtc tcatcgtgcg 240
tggtctctga tcttccatca cctgaccttc tccaggactg tcttctcacc ctccccact 300
ccctggtccc cgggagcagc tcttcttgcc cgactcactc acagtgcagg gaaaggaggc 360
agggaaaaga ccaggattct gtgagttctg aggttgccac acacaaagaa gctgtggttt 420
ctctgcctcg gccactgatg agactaaaac tggttccccc ttggagacgg cagatttcag 480
gctgatccct gcttaagccc tctcatcccc acgctggtcc tggattgat acaagaccca 540
gctggtgaca aagcctccaa tctggggggt ccacgagcct gggcctgaca ttccagaac 600
taccgccagt ggcgccaggc cccacagtc tgtggccgtg gtcttagccc ccagttccac 660
tctggatggg cctgtgacac cccaaagaga agaaggggac tctggatagg gtccccacat 720
ccagggcgtg gggagacat tggcatttgg gaaccatttt ccttcgaacg gcttccccctt 780
gagctgagca ttctgcttgc tgcagtagac gggtcgctt ttgccatac cgaaattttc 840
tgaaattaaa tcgcacaccc ccaccatttc ctctccctgg ggatctggag gaacatcata 900
catagtagggt gaatcgtttt gtagagtga gaatgcta atgtaaagcaaa tagtcaccca 960
cgttccttgt aaatccaaat gtttctatat tgtagctttg cttaaaatgg gggtcggccc 1020
caactgcata ctctctttg gcgggctggg gagcgggccc cagccgggac gggagggcag 1080
cgaccccgag gctcgtgga cgtgggagag agtggtgtgg gaagtcttga gcggaggagg 1140
ggatctgccc ttctccactc ctctcttggg ttcgcctcgg tttcctgtnc cccaccacc 1200
cgctgcccc gcggaagacc gccagtgag ccagccccc ccttccaggc gccttcgccc 1260

```

```

tggggatcca accaacttgt atcgagtggg cggggcaacg gctcccccatt ttccccgagc 1320
cccgcccaca gagctcttag ccaatcctat gcagagagca tctcctggca ggggtctctt 1380
cccaaccaga cccccccag gcacattagc gaccaggctt gggcttcccc agcgccccac 1440
caccaccacg tgcaggtgga gctctgggat gctatgttgg ggcggcaagc ggtgggcccga 1500
gggcccggta ggctagcacg ggaggttaagg gtggtatggg atggggcggg ggcggtctag 1560
ggcaatagga gagcagagaa tgggggaact tgagggtggg gggagggcac cggagccttg 1620
ccaccatccc aggacttttg gcaagtcacc cgcactccct gggcctcggg ttccccatct 1680
gtaaaatgat ggttaataata cttcacctac ctcatagggg aggttgtgag gccaccatca 1740
cctgacctgg ggtcaaggc aggaggactc cgaagggtgct acccgtgagc aaagtgtaat 1800
taccgaatcc tgactgcaag gccacctgc ccctccccc cagagcctcc agagctagct 1860
gaggccaacg caggcccatc cgtctcttca ctctgtcgca ggccttttca tgggcttcgt 1920
ctgccatctt tgtgggtgcc ctgacttag tccttatctt gtccctgggtt cctttcttgt 1980
gaccatctcc ccatgaaagt gctgtacaaa ttccaccgcg ccaggagacc ccgcacctgc 2040
cctctggcac cagatgccag ggaagggaca gaggaaaaca gccacaaaaca agccaggggg 2100
gctccccgga gccccagggg tggggatttg tggccactgt ttgtatgttc ttgagtgcaa 2160
gtgttttata aaaaaataaaa caaaaaccca ccatcacaaa aaaaaaaatt ttgcagcgaa 2220
gagaaatgaa gaaaaactga agaaaaaaa aaaacnggaa ataagaacca taaaaattt 2280
ttccaccaca cataccctct aagccagcaa gatttcctct ttgcaaaatc atatttttgt 2340
gggaatgggc cctgcttttt gtggcaaggc ctgttctgat taataaagga tcgtgaanan 2400
anaagttaaa nmtgtgattt caanaa 2426

```

&lt;210&gt; 8

&lt;211&gt; 510

&lt;212&gt; DNA

&lt;213&gt; Homo sapiens

&lt;220&gt;

&lt;221&gt; misc\_feature

&lt;223&gt; Incyte clone 3458076

&lt;400&gt; 8

```

gattttaaag cgcgcgctg ggcgcgctgg ggggcaagga agggggggcg gaaccagcct 60
gcncgcgctg gctccgggtg acagccgcgc gcctcgcca ggatctgagt gatgagacgt 120
gtccccactg aggtgcccc aagcagcagg tgttgagcat gggctgagaa gctggaccgg 180
caccaaaggg ctggcagaaa tgggcgctg gctgattcct aggcagtttg cggcagcaag 240
gaggagaggc cgcagcttct ggagcagagc cgagacgaag cagttctgga gtgcctgaac 300
ggccccctga gccctaccg cctggccac tatgggtccag aggctgtggg tgagccgct 360
gctgcggcac gggaaagccc agctcttgct ggtcaacctg ctaacctttg gcctggaggt 420
gtgtttggcc gcagattcac ctatgtgccg cctctgtgct ggaatggggg tagaggagaa 480
gttcatgacc atggtgctgg gatttggtcc 510

```

&lt;210&gt; 9

&lt;211&gt; 231

&lt;212&gt; PRT

&lt;213&gt; Homo sapiens

&lt;220&gt;

&lt;221&gt; misc\_feature

&lt;223&gt; Incyte clone 1864683

&lt;400&gt; 9

```

Met Gly Ser Leu Gly Leu Phe Leu Gln Cys Ala Ile Ser Leu Val
 1             5             10            15
Phe Ser Leu Val Met Asp Arg Leu Val Gln Arg Phe Gly Thr Arg
                20                25                30

```

```

Ala Val Tyr Leu Ala Ser Val Ala Ala Phe Pro Val Ala Ala Gly
      35              40              45
Ala Thr Cys Leu Ser His Ser Val Ala Val Val Thr Ala Ser Ala
      50              55              60
Ala Leu Thr Gly Phe Thr Phe Ser Ala Leu Gln Ile Leu Pro Tyr
      65              70              75
Thr Leu Ala Ser Leu Tyr His Arg Glu Lys Gln Val Phe Leu Pro
      80              85              90
Lys Tyr Arg Gly Asp Thr Gly Gly Ala Ser Ser Glu Asp Ser Leu
      95              100             105
Met Thr Ser Phe Leu Pro Gly Pro Lys Pro Gly Ala Pro Phe Pro
      110             115             120
Asn Gly His Val Gly Ala Gly Gly Ser Gly Leu Leu Pro Pro Pro
      125             130             135
Pro Ala Leu Cys Gly Ala Ser Ala Cys Asp Val Ser Val Arg Val
      140             145             150
Val Val Gly Glu Pro Thr Glu Ala Arg Val Val Pro Gly Arg Gly
      155             160             165
Ile Cys Leu Asp Leu Ala Ile Leu Asp Ser Ala Phe Leu Leu Ser
      170             175             180
Gln Val Ala Pro Ser Leu Phe Met Gly Ser Ile Val Gln Leu Ser
      185             190             195
Gln Ser Val Thr Ala Tyr Met Val Ser Ala Ala Ala Leu Gly Leu
      200             205             210
Val Ala Ile Tyr Phe Ala Thr Gln Val Val Phe Asp Lys Ser Asp
      215             220             225
Leu Ala Lys Tyr Ser Ala
      230

```

```

<210> 10
<211> 162
<212> PRT
<213> Homo sapiens

```

```

<220>
<221> misc_feature
<223> Incyte clone 3458076

```

```

<400> 10
Met Val Met Asn Phe Ser Ser Thr Pro Ile Pro Ala Gln Arg Arg
  1              5              10              15
His Ile Gly Glu Ser Ala Ala Lys His Thr Ser Arg Pro Lys Val
      20              25              30
Ser Arg Leu Thr Ser Lys Ser Trp Ala Phe Arg Cys Arg Ser Arg
      35              40              45
Arg Leu Thr His Ser Leu Trp Thr Ile Val Gly Gln Ala Gly Arg
      50              55              60
Ala Gln Gly Ala Val Gln Ala Leu Gln Asn Cys Phe Val Ser Ala
      65              70              75
Leu Leu Gln Lys Leu Arg Pro Leu Leu Leu Ala Ala Ala Asn Cys
      80              85              90
Leu Gly Ile Ser Gln Ala Pro Ile Ser Ala Ser Pro Leu Val Pro
      95              100             105
Val Gln Leu Leu Ser Pro Cys Ser Thr Pro Ala Ala Val Gly His

```

	110		115		120
Leu Ser Gly Asp Thr Ser His His Ser Asp Pro Gly Arg Gly Ala					
	125		130		135
Arg Leu Ser Pro Gly Ala Ser Ala Xaa Arg Leu Val Pro Pro Pro					
	140		145		150
Leu Pro Cys Pro Pro Arg Ala Pro Ala Gly Gly Phe					
	155		160		